Strategies to Detect Genetic Diversity in Plants

Dissertation

zur Erlangung des akademischen Grades

„doctor rerum naturalium“ (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät

der Friedrich-Schiller-Universität Jena

von Diplom-Bioinformatiker Thomas Schmutzer

geboren am 24.06.1982 in Leipzig

Gutachter:

1. PD Dr. Matthias Platzer
   Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut Jena, Deutschland.
2. Prof. Dr. Günter Theißen
   Friedrich-Schiller-Universität Jena, Deutschland.
3. Prof. Dr. Bernd Weisshaar
   Universität Bielefeld, Deutschland.

Termin der Verteidigung:

04.04.2016

Contents

List of figures

List of tables

List of abbreviations

AFLP	Amplified fragment length polymorphism
BAM	Binary compression format of SAM format
bp	Base pair
BWA	Burrows-Wheeler aligner
CAP-seq	Capture sequencing
CBD	Convention of Biological Diversity
cDNA	Complementary DNA
CDS	Coding sequence
CNV	Copy number variation
CVC	Combinatorial variant calling
DArT	Diversity array technology
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
ddNTP	Dideoxynucleoside triphosphate
EST	Expressed sequence tag
FAO	Food and Agriculture Organization of the United Nations
FISH	Fluorescence *in situ* hybridization
Gbp	Giga base pair
GBS	Genotyping by sequencing
GCR	Genome complexity reduction
GS	Genomic selection
GWAS	Genome wide association study
IBSC	International Barley Genome Sequencing Consortium
IHGSC	International Human Genome Sequencing Consortium
INDEL	Insertion or deletion polymorphism
IWGSC	International Wheat Genome Sequencing Consortium
KASP	Kompetitive allele specific PCR
Kbp	Kilo base pair

MAF	Minor allele frequency
MAS	Marker assisted selection
Mbp	Mega base pair
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
NQS	Neighboring quality standard
ORF	Open reading frame
PAV	Presence/absence variation
PCR	Polymerase chain reaction
PIC	Polymorphic information content
QTL	Quantitative trait locus
RAD	Restriction site associated DNA
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
SAM	Sequence alignment/map format
SBS	Sequencing by synthesis
SFF	Standard flowgram format
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
SSR	Simple sequence repeat
SV	Structural variation
TGS	Third generation sequencing
TREP	Triticeae Repeat Sequence Database
UTR	Untranslated region
VCF	Variant call format
VP	Variant position
WGS	Whole genome shotgun
WGS-seq	WGS sequencing

“I am thrilled to see my genome.” James D. Watson

1. Introduction

The term '*in silico* SNP discovery' was first introduced in the year 1999 by Picoult-Newberg et al. [1]. At that time, the rapid development of sequencing technologies together with its deeply embedded applications in today's science was a vague premonition. Nevertheless, that statement became remarkably correct and already introduced the objectives of this thesis.

The year 2010 was named the "International Year of Biodiversity" by the Convention of Biological Diversity (CBD, http://www.cbd.int). The objectives of this convention were formulated as three major issues: 1. "Conservation of biological diversity", 2. "Sustainable use of its components", and 3. A "fair and equitable sharing of benefits arising from genetic resources". With these strategic decisions and recommendations, a sustainable development plan has been promoted to achieve the well-being of both humanity as well as the natural world since the first meeting in 1994 [2]. In addition to vital government support and commitments to achieve the goals of the CBD, the discovery of new genetic variation is a central demand. In section [4.1 and 4.2] of this thesis, I will review biotechnological as well as computational methods applicable to detect novel DNA variation in plants. Furthermore, accuracy of variant prediction is an essential need to assess the diversity of a species. Measurements that support this requirement and that assist to score and discard less valid predictions would be beneficial. As a consequence, I developed two methods to improve the accuracy of diversity prediction. These are captured and discussed in section [4.3] of this thesis. The conversion of the determined diversity information into application is discussed in section [4.4].

Genome, exome and transcriptome sequencing provide access to the genetic information of virtually any species at relatively low cost. Depending on which of these strategies is applied to investigate a complete or reduced genome the costs and computational requirements differ strongly. But concurrently, research insights and their applications differ, too. Besides the pure deciphering of the DNA sequence for a species, the availability of genomic data furthermore opens the possibility to reveal the genetic diversity between different species. This information is contained in these genomic data as well. The integration of DNA markers has revolutionized plant science research in terms of pace and precision. A crucial role is the discovery of reliable variants that are applicable for subsequent analysis. The objectives of this thesis are to enlighten the applicability of different genomic strategies towards a successful and accurate investigation of diversity.

1.1. Survey of genomic resources for rye, maize and barley

Three cereal crops are analyzed in the framework of this thesis. Rye (*Secale cereale* L.), maize (*Zea mays* L.) and barley (*Hordeum vulgare* L.) are each characterized by an extraordinary complexity because of their large genome size and high repeat content.

If the ordering of the studied crops should reflect the worldwide economic importance one would start with maize as the most economically relevant crop, followed by the two closely related species barley and rye. Related to this worldwide economic importance the availability of genomics resources has been and still is very different as illustrated in the overview of Figure 1. For maize, in the year 2010, a fully annotated reference sequence and a large set of markers existed to support research initiatives [3]. For barley, a comprehensive collection of diverse resources was available in 2010 including DNA markers [4,5] and genomic sequences [6–8]. In the International Barley Genome Sequencing Consortium (IBSC), that was founded 2006, scientists combined research initiatives to establish a genome reference resource [9]. For rye, in 2010, almost no sequence information was available with a small exception of 10,000 expressed sequence tags (EST) and an initial survey sequencing of the short arm of chromosome 1R providing 2,778 BAC end sequences totaling 2 Mbp [10]. Therefore, a German consortium of scientists was formed to accomplish important steps for the minor crop rye. The construction of resources comparable to the established ones in maize are the challenges and long-term objectives in rye and barley genomics. The common aim of all these initiatives is to provide a draft reference sequence. Draft genome sequences have a lower accuracy than finished sequences (e.g. segments are missing or have the wrong orientation or order), but have a large value for a variety of research studies because most genes are already represented [11]. Having in hand such a resource improves the ability to assess new perspectives for breeding and facilitate new approaches in crop research [12].

1.2. DNA sequencing

The first generation of DNA sequencing method was initiated in the year 1975 by Frederick Sanger, who invented the chain-terminating method for DNA sequencing. The method allowed to determine the sequence of nucleobases along a DNA polynucleotide chain. The building units of the genomic DNA, the nucleotides, contain one of the four nucleobases (A, C, G or T), being either a purine (A and G) or a pyrimidine (C and T). The initial publication in 1975 was a milestone in molecular genetics [13]. Two years later Maxam and Gilbert [14] published their protocol that rapidly became a widely used sequencing method, since it was able to use purified double-stranded DNA instead of cloned single-stranded DNA. However, with a simultaneous publication in 1977, Sanger

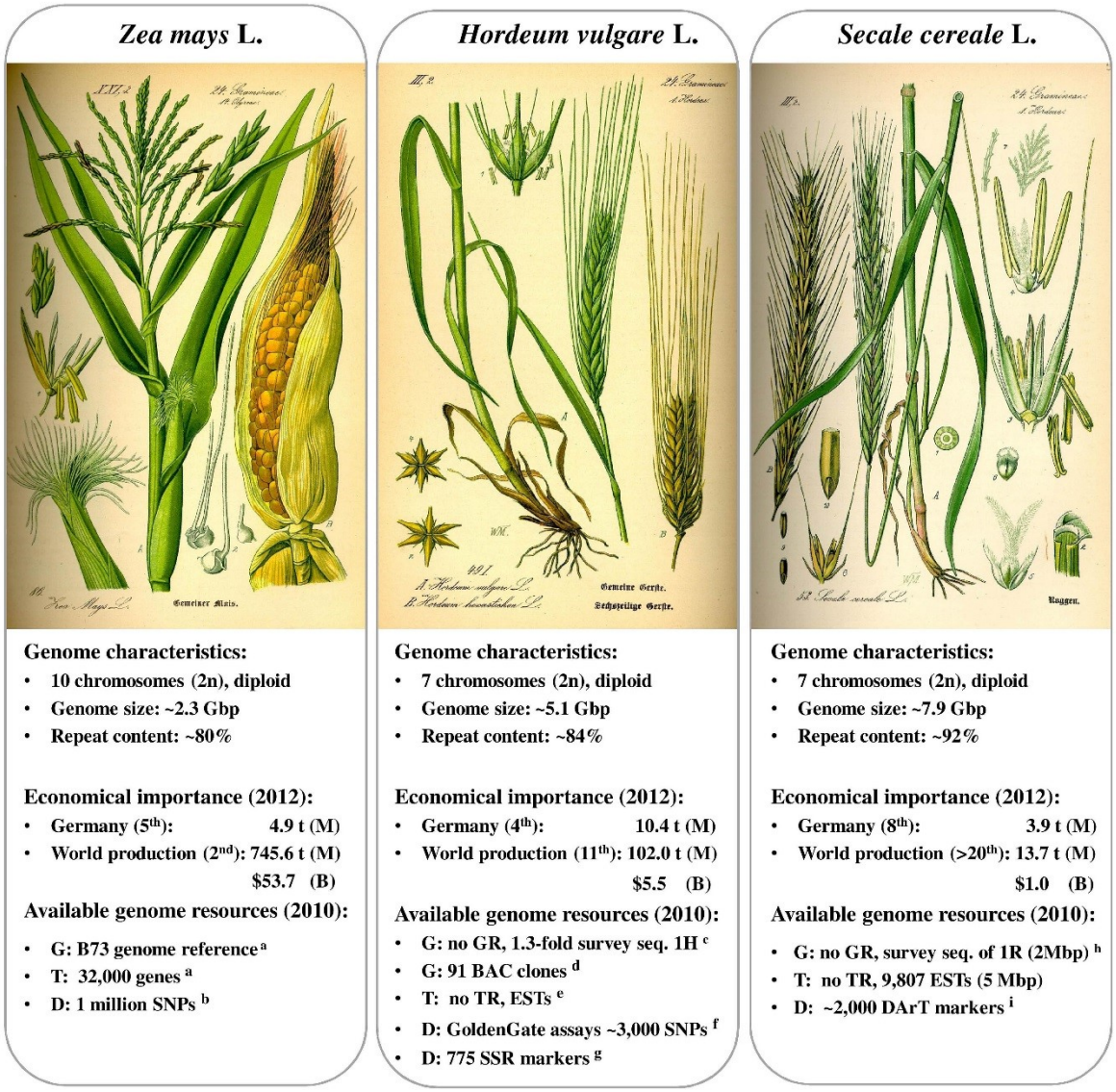

Figure 1 Survey of crop plant characteristics. Botanical drawings of plants were adapted from [277]**.** Agronomical values are based on reports of the FAO (2012). Values of production quantity are in given in millions of tons and billons of dollar. For Germany the statistics of FAOSTAT were analyzed [278]. For world production the sum of the top twenty countries with highest the productivity is used. The overview of the availability of genomic resources is based on publications that were released in or before the year 2010, selected with the aim to refer to the most advanced genomic resources available at that time. Due to different individual research focuses the given perspective is rather a subjective selection and does not claim a complete objective overview. Given values refer to (G)enome, (T)ranscriptome and (D)iversity for different resource types. Because of space limitation the abbreviations for genome reference (GR) and transcriptome reference (TR) are used in the graphic. Values are taken from **a)** Schnable et al. [3]; **b)** Lai et al. [279]; **c)** Mayer et al. [8]; **d)** Steuernagel et al. [7]; **e)** Stein et al. [6]; **f)** Close et al. [5]; **g)** Varshney et al. [4]; **h)** Bartos et al. [10] and **i)** Bolibok-Bragoszewska et al. [280].

provided a less complex and better scalable method that was able to be used in sequencing kits. Thus, it became the method of choice for the upcoming 25 years [15]. This method applies the sequencing by synthesis (SBS) concepts, using single-strand DNA as sequencing templates. Together with a DNA primer and deoxynucleoside triphosphates (dNTP), the DNA polymerase elongates the new DNA fragment. This replication process is repeated until a dideoxynucleotide triphosphate (ddNTP) terminates the reaction. The product of a sequencing reaction is individually loaded into one lane of a gel and the DNA fragments are subsequently size separated by gel electrophoreses. This allows for the determination of the order of nucleotide bases for a DNA sequence, based on the varying length of different synthesized DNA fragments. Various technical inventions like the use of primers tagged with a fluorescence dye, which allowed sequencing in optical systems [16], have continuously decreased cost and increased throughput. But in general, the core of the technology to decipher genes and genomes did not change.

The International Human Genome Sequencing Consortium (IHGSC) applied Sanger DNA sequencing for deciphering of the human genome, which has a size of over 3 billion nucleotides [17,18]. The achievements of this breakthrough project have shown the requirement for high throughput and cost-efficient sequencing. Therefore, a funding program was initiated by the National Human Genome Research Institute (NHGRI) with the aim to reduce DNA sequencing costs by four orders of magnitude within the next ten years [19]. This gave rise to the development of next-generation sequencing technologies.

1.2.1. Next generation sequencing technologies

In the last decade, DNA sequencing experienced an enormous technological shift leading to a massive increase of throughput and a sharp decline of costs per-base (Figure 2). The corresponding DNA sequencing platforms are termed second or next generation sequencing (NGS), to express the advance (e.g. massive parallelization) that was achieved compared to the first generation of DNA sequencing (Sanger). Current DNA sequencing technologies decipher the order of nucleotide bases only within short regions (100-20.000 bp) that are referred to as ‘reads’. To reconstruct the entire sequence of usually much longer DNA/RNA macromolecules requires the subsequent process called assembly that aims to generate a larger continuous sequence (‘contig’) from the reads.

The two sequencing platforms Roche 454 (www.454.com) and Illumina (www.Illumina.com) have been used in projects that are embedded in this thesis and therefore, will be briefly introduced in this section. Several other NGS technologies exist, e.g. SOLiD (http://www.lifetechnologies.com), Ion Torrent (https://www.thermofisher.com/de/de/home/brands/ion-torrent.html) and Complete Genomics (http://http://www.completegenomics.com/). Moreover, a third generation of sequencing (TGS) platforms began to emerge, with PacBio (http://www.pacificbiosciences.com) as the only one

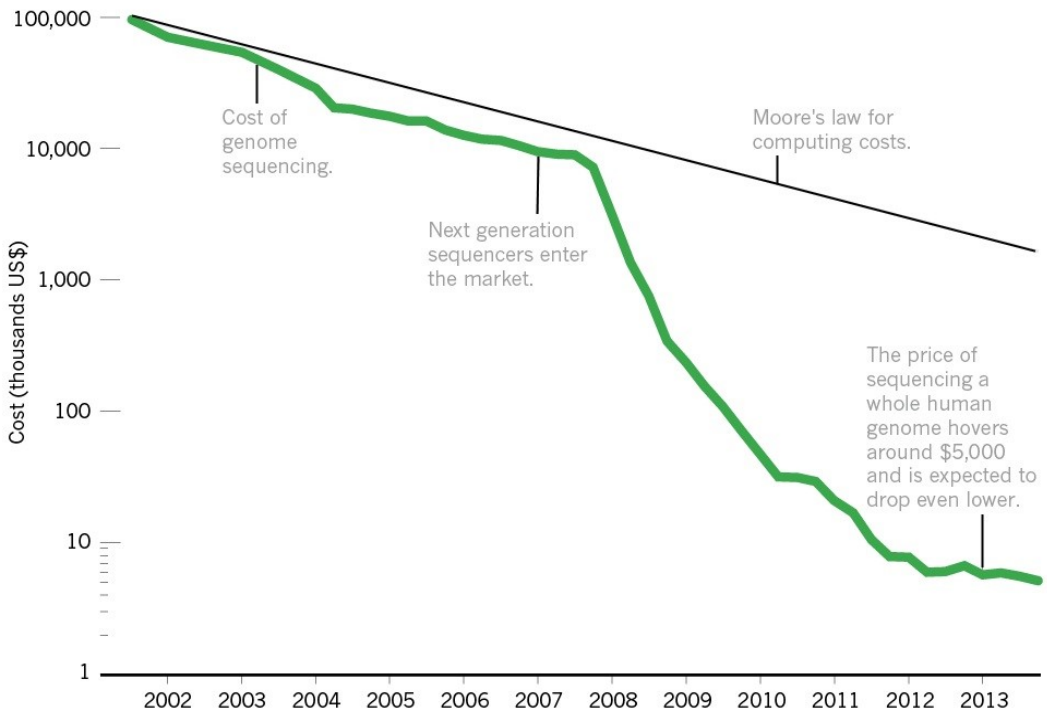

Figure 2 Decline of DNA sequencing costs. Even after the Human Genome Project was declared to be finished in April 2003, the sequencing cost still followed Moore's law. It is predicting a decline of computing cost in exponential scale. Until 2008, shortly after NGS technologies were established, the sharp and continuous decline of costs started. Figure is adapted from [281].

currently commercially available. With the Personal Genome Machine by Ion Torrent [20,21] or the MiniION developed by Oxford Nanopore (https://www.nanoporetech.com/) other TGS platforms are developing. In contrast to previous generations, these technologies aim at single-molecule sequencing. For a broader introduction and comparison several reviews have been published [22,23].

The technology of 454 sequencing became commercially available in 2005 with the company 454 Life Science, which was acquired by Roche in 2008 [24]. The technique applies a SBS method published as pyrosequencing [25,26]. Roche 454 sequencing runs in parallel on a large-scale and is capable to produce up to 700 Mbp of sequence data in a 10-hour run [27]. It produces sequence data that exceed in more than 85% of the reads a length of 500 base pairs (bp), having on average 700 bp [28]. Current instruments are the GS FLX+ system and the GS Junior system. At the end of the DNA sequencing process a standard flowgram format (SFF) is produced, containing the DNA sequencing results.

The second provider of NGS platforms was the company Illumina. The Illumina NGS approach was established in 1997 by S. Balasubramanian and D. Klenerman. They invented the approach using a solid phase sequencing in combination with reversible dye-terminators. With a successful venture funding they formed the company Solexa in 1998, who launched the sequencing technology in 2006.

Table 1 Overview of DNA sequencing technologies from 1rst to 3rd generation. Listed DNA sequencing platforms include the two sequencing platforms (454 & Illumina) used to generate results for this thesis [2.1, 2.2 and 2.4]. The list also includes Sanger technology, representing the first generation of DNA sequencing, and PacBio representing the third generation. Technical descriptions are extracted from vendors specifications (accessed 20th of October 2014). The ranking (cost per Mbp) is adapted from the comparison published by Shendure et al. [171].

Sequencing system	ABI 3730xl6	GS FLX [1,2,3]	HiSeq 2000 (Hiseq 2500)[1,2,4]	PacBio RS II[1,2,4]
Generation	1rst	2nd	2nd	3rd
Vendor	Life Technologies[5]	Roche 454[6]	Illumina[7]	Pacific BioScience[8]
Amplification	clonal	emulsion PCR	bridge PCR	no amplification
Data output per run	0.9 Mbp	700 Mbp	200 Gbp (600 Gbp)	400 Mbp
Time per run	2 hrs	23 hrs	8 days (10d)	20 -180 minutes
Cost per run	$50	$6,000	$20,000	$400
Cost per Mbp[2]	very high (>$500)	high (>$1)	low (<$0.1) (very low ~$0.05)	high (>$1)
Cost per instrument	$90,000	$500,000	$650,000 ($740,000)	$700,000
Accuracy	99.999%	99.9%	99.9%	99.999%[9]
First error	failed detection of heterozygous INDELs	INDELs	mismatch error	random error type
Read length[2]	long (~500 bp)	long (~700bp)	short (~100 bp) (short (~125 - 150 bp))	very long (~1,000 bp)

[1] http://allseq.com
[2] Shendure and Lieberman Aiden (2012)
[3] Liu et al. (2012)
[4] Quail et al. (2012)
[5] http://www3.appliedbiosystems.com
[6] http://www.454.com
[7] http://www.illumina.com
[8] http://www.pacificbiosciences.com
[9] The accuracy of PacBio sequencing is refering to the vendors specification requiring a 30-fold coverage. Quail et al. (2012) reported single pass accuray of <90% (QV10).

In 2007, Solexa was acquired by Illumina. The sequencing process is described in more detail by several reviews [29,30]. The enormous throughput of Illumina machines justifies its leading position among the different sequencing platforms. For example, the currently widely used sequencer Illumina HiSeq 2500 generates 120 Gbp in rapid run mode (27h) or up to 600 Gbp in the high output mode (10d) (Table 1). A typical run produces ~30 Gbp per lane (HiSeq 2500 has a total of 2x8 lanes). In addition, Illumina provides an alternative system, the MiSeq. This produces less output, but has an immense improvement of read length (maximal output setting is 2x300 bp in a 65 hour run, producing 15 Gbp of data output), compared to the 2x100bp on a HiSeq 2500 high output mode. Several sample preparation kits are available to support the preparation process and offer highly customizable solutions. Because of the extremely high-throughput and the broad scope of support, Illumina currently is the leading vendor of sequencing technologies [23].

Furthermore, different TGS technologies currently are reaching the phase of consolidation, working on the reduction of systematic errors (e.g. error profiles) or the improvement of technical difficulties (e.g. deletion errors at homopolymers) [31,32]. Referring to Schadt et al. [33] the TGS technologies seek to improve on six different levels in comparison to the second generation (throughput, time, read length, accuracy, DNA quantity and costs). However, at the current state most of these challenging goals require a deeper refinement [34], where the increase of nucleotide accuracy is one

of the most demanding needs. Nevertheless, among different existing TGS platforms the Pacific Bioscience (PacBio) has been applied successful in various projects [35,36].

Compared to the cost of Sanger sequencing, that rarely fell below $500 per Mbp [37], the cost for DNA sequencing was tremendously decreased with NGS (Figure 2). The decline successfully reached the four orders of magnitude, which were set as challenging goal by the NHGRI ten years ago [19]. This shows that sequencing will have a constant crucial influence on goals and perspectives of research. As described in detail by van Dijk et al. [38], the NGS technology has significantly driven the change of research and its improvements during the last ten years, and it will seamlessly pave the way for new perspectives in research studies. As intended by the FAO, the focus on biodiversity will challenge science to ensure nutrition of the world population in future [39,40]. In addition to this enormous challenge, the questions of personal genomics and clinical diagnostics are expected to be the driving forces for upcoming technological developments [41].

1.2.2. DNA sequencing errors

The quality achieved with high-throughput sequencing technologies is, in its current technological status, lower than with traditional capillary sequencing (Sanger) regardless of the type of NGS platform. As emphasized in the section introducing DNA sequencing technologies, each platform has its inherent error models (Table 1). While the precise investigation of sources for putative errors constitutes a research topic on its own, this section aims to give a brief description of the most common errors observed. A detailed perspective on sequencing technologies and error profiles is given by Kircher and Kelso [37] as well as Harismendy et al. [42].

All technologies of the first and second generation of sequencing used amplification steps to increase the number of DNA templates and thus the signal intensity. In consequence, PCR-based duplication is one of the major error sources resulting in identical sequence reads, which require correction and decrease the effective sequencing output [43]. It should be noted that the single-molecule sequencing utilized by TGS technologies has the advantage that this source of an error is eliminated [44]. However, in the framework of this thesis all analyzed sequence data sets were generated using the 454 or Illumina sequencing technologies, thus requiring pre-processing steps to correct duplicated reads.

In addition to the duplicated read error that originates from the DNA preparation process, the pyrosequencing approach, utilized by 454, has another important drawback. The technology is error prone when calling homopolymer sequences (long stretches of identical nucleotides). In the

sequencing-by-synthesis process the emitted signal strength is proportional to the length of the homopolymer [26]. However, saturation is reached at a stretch of eight consecutive nucleotides. Because of this limitation, the major sequencing errors observed are insertions and deletions (INDEL) that are significantly more frequent in 454 (median error rate is 0.3% for deletions and 0.2% for insertions) than in Illumina data (average 0.005%). This discrepancy is strictly limited to homopolymer stretches, while similar error profiles for both sequencing platforms are observed beyond these critical regions [45]. Referring to the publication of Harismendy et al. [42], INDELs have also been discovered as sources of errors in traditional Sanger sequencing although on a much lower scale and with respect to heterozygous positions.
Substitutions ('mismatches') are the most frequent type of error observed in Illumina sequences [30,46]. The global substitution error rate has been revealed to be about 0.16%, estimated in a plant sample control data set [47]. Several other types of sequence errors, e.g. underrepresentation of AT-rich and GC-rich regions, are associated with both Illumina and 454 sequencing when whole-genome shotgun sequencing is used [42]. An amplification bias has been identified as source of this problem and methods have been proposed to apply correction models [48].

With the increasing amount of produced sequence data, several tools have been developed to use these massive amounts of read data to correct intrinsic sequencing errors. Referring to a detailed review of Yang et al. [49] the different approaches can be classified into three categories. All approaches have the use of efficient data structures based on *k*-mers (oligomers of length *k*) in common. Applied methods use these short motifs to construct efficient data structures like suffix arrays or enhanced suffix arrays [50]. In the first method of 'multiple sequence alignment', reads are screened to identify related subsets of reads that share identical *k*-mers and subsequently a correction is performed within these clustered subsets. The second method of '*k*-spectrum' uses all *k*-mers, which are present in decomposed reads to correct less reliable reads (low *k*-mer frequency) with minimal edit operations. In the third category, suffix tree or suffix array structures are used to correct putative sequence errors by identifying rare substrings [51]. Yang et al. [49] revealed a general requirement of error correction to increase and validate raw sequence data. At the same time, they emphasize the need for further algorithmic improvements to achieve a reliable correction of the broad spectra of failures.

1.2.3. Scope of sequencing

Next generation sequencing has great potential and is offering the unprecedented opportunity to assess also complex plant genomes [52]. In the year 2000, the first complete genome of the model plant *Arabidopsis thaliana* became available [53]. In the year 2013, the genome sequence of *Aegilops*

tauschii [54] was published and with it the 50th plant genome sequence [55]. Genome sizes of these 50 species range between 82 Mbp of the bladderwort (*Utricularia gibba*) to 19.6 Gbp of the Norway spruce (*Picea abies*). With a genome sequence that is 100 times larger than the one of *A. thaliana* [56], particularly the latter clearly indicates that these projects have been made possible by the advances in NGS [55]. Ploidy, as well as heterozygosity or gene duplications are just a few of the challenges that complicate the understanding of plant species. Handling the high amount of sequence data is just the beginning of the computational challenges. Very different kinds of questions can be addressed to these data resources and hence, a continuous development and improvement of algorithms is required. This thesis aims to improve the aspect of diversity calling.

With the dramatic reduction of sequencing costs, the analysis of diversity on a large scale has become feasible for many species. Nowadays, sequencing projects are underway for even highly repetitive, large and complex genomes like barley (5 Gbp) and rye (8 Gbp). Furthermore, whole population studies are in progress like the 1001 Genome project, aiming to sequence 1,001 strains of the reference model plant *A. thaliana* [57]. The expanding scope of sequencing together with the still decreasing costs will enable the research community to consider novel scientific questions. A main objective of DNA sequencing is still to determine eventually the complete nucleotide sequence of a genome. Therefore, whole genome shotgun (WGS) projects are considered for various species to enable functional genomics and gain deeper understanding of plant genome structure, evolution and complexity [56,58,59]. However, sequencing will have an expanding scope of further applications that are more feature-oriented. One exemplary method is Methyl-seq for the discovery of DNA methylation sites in the genome [60,61] that provide novel insights into genetic regulation. Other aspiring possibilities are ChIP-seq methods, that investigate the protein-DNA interactions using immunoprecipitation [62,63], or Hi-C, which reveals the three-dimensional structure within the genome by interaction and contact of genomic DNA [64]. This list of applications can be extended by two methods that were used in the projects integrated in this thesis [2.1 and 2.2]. RNA-seq and CAP-seq (also known as ‘capture sequencing’) are sequencing strategies that perform a genome complexity reduction (GCR). With this, a more focused perspective towards the transcriptome or a particular selection of genome segments (e.g. genes) is achieved. The next section provides an overview of these two DNA sequencing strategies applying GCR and the WGS sequencing without complexity reduction.

1.2.4. DNA sequencing strategies for the detection of diversity

NGS paved the way to access literally any genome of interest by WGS sequencing. In addition, different GCR methods can be applied to assess the DNA sequence of a species and to study its

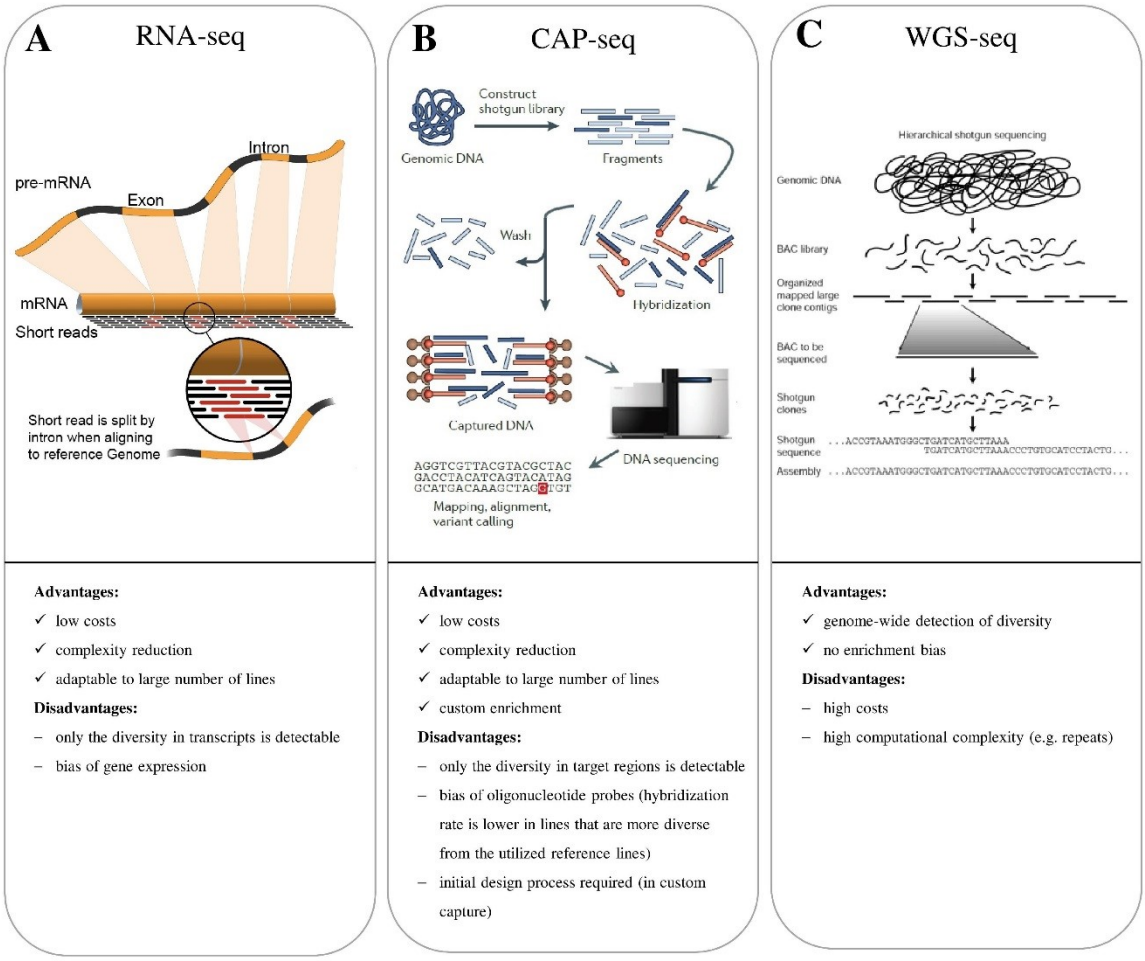

Figure 3 Overview of applied DNA sequencing strategies. Each sequencing strategy has characteristics that provide advantages or disadvantages. This information is provided together with an illustrated outline of the workflow/concepts. **A)** RNA-seq represents the coding part (exons) of genes. Figure adapted from [65]. **B)** CAP-seq approach targeting custom sequence regions. Figure adapted from [282]. **C)** WGS-seq approach represents the complete genomic sequences of a species. Figure adapted from [18].

genetic diversity [43]. This thesis will illustrate how RNA-seq, CAP-seq and WGS-seq can be implemented in the framework of diversity studies and what benefits each method provides. In addition to these methods, several array technologies like Affymetrix (http://www.affymetrix.com) or NimbleGen (http://www.nimblegen.com) exist for genotyping. It is important to emphasize that these assay-based genotyping methods represent only a small portion of DNA markers and thus only a small fraction of the complete genetic diversity. In contrast, DNA sequencing enables the *de novo* detection of SNVs to reveal novel variations within a genome. The perspective of this thesis mainly focuses on the discussion of the three DNA sequencing methods. In comparison to an assay-based diversity study they all have the advantage of capturing the genetic diversity in a much more systematic and thorough way.

Transcriptome sequencing (RNA-seq) has evolved as a method to study gene expression [65,66]. As proven by Zhao et al. [67], the method is very accurate for transcriptome profiling in comparison to microarray technology. Both sequencing platforms, 454 and Illumina, offer RNA-seq as a powerful tool to access the expressed transcripts of a species. The method is capable of analyzing gene expression at the level of transcripts and splicing variants, although this would require additional replicate sequencing and a solid experimental study design [68]. RNA-seq uses complementary DNA (cDNA) libraries for sequencing [69,70]. This cDNA is synthesized from the input RNA by the enzyme reverse transcriptase. The input RNA (mRNA) represents the coding sequence of a gene (exons), excluding the non-coding intervening sequences (introns) [71]. If strand-specific RNA-seq is utilized one possible challenge are artifacts that can appear by spurious synthesis of second-strand cDNAs during the reverse transcription reactions [72]. Particularly in studies, which aim for the analysis of strand-specificity transcription profiles, these antisense artifact need to be excluded. Another possible error in RNA-seq studies are artifacts that result from incompletely spliced pre-mRNA, which can contribute to the construction of spurious isoforms [73]. This is of particular interest when total cellular RNA (mixture of nuclear and cytoplasmic RNA) is used to construct *de novo* transcript assemblies. For a broad perspective of the advantages as well as challenges of RNA-seq methods Ozsolak and Milos [71] provide a detailed overview.

However, beside the expression profiling, variation calling on the transcribed proportion of a genome offers another cost effective approach to study the genetic diversity of a species. Therefore, several individuals of a species have to be selected that reflect the broad spectrum of diversity within the species. The significant reduction in complexity is very beneficial for sequencing large and repetitive genomes (Figure 3A). 454 sequencing technology has been validated for its accuracy in SNV detection in several species like maize [74] and also in hexaploid oat [75]. Nevertheless, application in the complex genome of rye had to be analyzed for reliability, which was the aim in the first publication included in this thesis.

In contrast to transcriptome sequencing, the CAP-seq method can address literally any part of the genome sequence. One of the main system providers of sequence capture methods for NGS are Roche NimbleGen (http://www.nimblegen.com/seqcapez) and Agilent (http://www.genomics.agilent.com) with its SureSelect platform. Initially in the sequence capture design target regions are selected, which subsequently are used to design custom specific oligos that will be ligated to an array surface (on-array capture) or alternatively are used for an in solution-based (in-solution capture) target enrichment [76]. The subsequent hybridization step with sample DNA allows to filter for target DNA, which is captured by the complementary oligo on the array (Figure 3B) or in-solution. After hybridization the target DNA is eluted and in the subsequent DNA sequencing process only reads

for the selected target regions are produced. A disadvantage of the capturing method is the hybridization that can be biased towards the reference lines that was used to construct the oligonucleotides. As a result, more diverse lines have a lower possibility to be captured. To avoid false negative results priority should be given to target regions with low 'missing data' [77]. Another type of target enrichment methods is exome capture that is targeting sequences corresponding to protein-coding exons [78]. Whole-exome target sequencing was adapted and has been extensively used in plants like barley [77], pine [79] and wheat [80] in the last few years. Target enrichment therefore provides a promising technology for variant discovery.

Sequencing a genome by whole-genome shotgun sequencing (WGS-seq) is the third DNA sequencing strategy investigated in this thesis for application of diversity discovery (Figure. 3C). With the advent of NGS technologies WGS sequencing became feasible in terms of sequencing costs also for complex genomes. Nevertheless, in species with large genomes the sequencing cost still is a limiting factor. The genome size of *Secale cereale* (7.9 Gbp) is estimated to be 64-times larger compared to the 125 Mbp *A. thaliana* genome (Arabidopsis Genome Initiative, 2000) and thus WGS sequencing in rye requires far more financial support in comparison to species with small genomes like sorghum with 730 Mbp [81]. However, the advantage of having access the complete genome sequence exceeds, in many research projects, the disadvantage of higher sequencing costs. Furthermore, the scope of questions which can be addressed to an established WGS reference genome are much more diverse. Sequencing a complete genome with NGS generates massive amounts of raw sequence data, particularly when large species collection are re-sequenced like rice [82] or species with large and repetitive genomes like pine [83,84] are studied. As a result, complex computational methods are required to reveal solid and useful insights. On the one side scientists are challenged with an increased complexity and on the other side they are attracted by the idea to have in hand a complete genome sequence.

1.3. *De novo* assembly and alignment of reads

A main goal of DNA sequencing is to determine the DNA sequence and thus to investigate the genomic architecture of a species. The aim of the assembly process is to reconstruct the entire sequence from reads and to generate larger contigs. Therefore, a consensus sequence is constructed that is defined as that sequence, which is most similar to all overlapping reads. The resulting reference sequence works as the standard sequence for a species, representing all major alleles (nucleotide bases) observed in the consensus sequence. Two major assembly concepts exists to reconstruct the DNA sequence from short sequence reads [85]. First category of assembly concepts is the overlap-layout-consensus assembly concept [86,87]. Through partial overlapping of reads, the consensus

sequence is elongated, constructing a reference sequence. The second concept of De Bruijn graphs is the most widely used method to assemble short sequence read data. It uses directed graphs to represent substrings of reads and reconstructs the sequence by traversing the graph in an Eulerian cycle [88,89]. Differences in sequence reads can have various sources. Principle aspects that need to be considered are for instance sequencing errors or differences based on nucleotide diversity. The construction of a consensus sequence without a reference is called *de novo* assembly, and this task remains challenging [90]. In re-sequencing projects with a known reference sequence another bioinformatics problem has to be addressed, called read alignment [91]. With this, the computational problem is to locate the correct position of a sequence read within the given reference backbone.

Raw sequence data are generated in the standard format FASTQ. Technically, sequencing machines produce technology specific raw data as the primary output of the DNA sequencing process, from which in the next step FASTQ files are derived. In the following I will use the term 'raw sequence data' referring to the unprocessed sequence data in FASTQ files. In the FASTQ format, nucleotides have assigned quality values that are associated with an ASCII-code, which corresponds to a PHRED score. This score gives the probability that a nucleotide is incorrect. It is a quality measurement that is logarithmically related to the base-calling error probability [92]. For example, a Phred quality score of 10 assigned to a nucleotide means that this base is incorrect in 1 of 10 cases (90% accuracy), where a Phred quality score of 30 would mean that this base is incorrect in 1 out of 1000 cases (99.9% accuracy).

1.4. Genome diversity

Biodiversity is divided into three subgroups: ecosystem diversity, species diversity and genetic diversity. The ecosystem diversity captures all different forms of habitats. Species diversity is represented by the number of species living in a particular ecosystem. The genetic diversity refers to the total of all the genetic variations within a species. These can occur on the level of chromosomes (e.g. duplications), segments of chromosome (e.g. translocations, deletion, segmental duplication) or on the level of nucleotides. The genetic diversity serves an important function for organisms to adapt to environmental changes [93]. Diverse populations have a higher adaptability for changing environmental conditions (e.g. drought or disease), because individuals might possess a trait offering a particular benefit (e.g. resistance) to overcome the risk of extinction. In agriculture, breeders take advantage of genomic variants to further improve existing lines and to develop advanced varieties. Genomic sequence resources provide valuable insights to study the genetic diversity on a genome-wide level, which will be the focus of this thesis. Changes in the genomic DNA are frequent [94]. Therefore, diversity studies at the nucleotide level have the benefit that genetic characteristics can be

observed frequently and throughout the complete genome as shown for many plant species like *Oryza sativa* [95,96], *A. thaliana* [57], or *Solanum lycopersicum* [97].

Most eukaryotes are diploid, meaning that they have two sets of homologous chromosomes. In this respect, zygosity is the state of sequence similarity of alleles at a particular locus on the DNA molecule [98]. If both alleles are identical, this position is referred as homozygous. If not, the locus is considered heterozygous. Other, more specific, forms of zygosity exists for diploid organism like hemizygous, describing that only a single copy of a gene is present, or nullizygous, where both copies of a gene are missing [99,100]. Throughout this thesis I will mainly focus on homozygous and heterozygous conditions.

The mating-system of the analyzed crops has to be mentioned because it introduces a fundamental difference. Barley is a self-fertilizing plant, whereas maize and rye are obligatory out-crossing species. In outcrossing species individuals transmit 50% of their genomic material to their offspring, whereas in species with strict self-fertilization 100% is transmitted [101]. Self-fertilization ('selfing') increases homozygosity and reduces the genetic diversity of a species [102]. On the other side, the consequence of outcrossing is a high intraspecific variation between individuals of species. In many plants like *Arabidopsis lyrata* [103] or rye [104] self-incompatibility system could be determined, where their evolution is expected as mechanism to prevent the plant from the negative consequences of inbreeding. However, several highly successful selfing species exists like wild barley [105], that emphasize the evolutionary benefit of this mating-system [106]. Finally, it should be mentioned that many plant species have a mixed mating-system, where occasional outcrossing events result in new heterozygous lines [107]. The main focus of this thesis is the reliable detection of diversity. In genomic studies, especially in those of non-model species with complex genomes, sequencing projects mainly use inbred lines, because these largely reduce the complexity of the study on several levels (e.g. *de novo* assembly, traits studies). Hence, I do not discuss and differentiate in closer detail the different mating-systems. For a more detailed description of mating-systems and their importance in plant breeding or establishment of germplasm collections I refer to several publications [102,106,108,109].

The preservation of diversity is a fundamental issue in natural habitats [110]. In an agricultural context, biodiversity is also seen to be necessary as a source for new breeding concepts [111]. Genebanks will play a key role to overcome one of the big challenges of this century. Nutrition and food security for a global human population, which is estimated to grow up to 9 billion people by 2050 [112], are only two of these big challenges. The Genebank of the Leibniz Institute of Plant

Genetics and Crop Plant Research in Gatersleben seeks to preserve this diversity of crop plants by collecting and conserving different genetic varieties. Climate change and the need to cultivate crops in less optimal environments are forcing breeding initiatives to develop new strategies and varieties. Solutions to improve plant tolerance against abiotic stresses, like drought and cold, and biological stresses, like plant disease, are required. Many wild species carry useful characteristics that are not present in cultivated elite lines used in breeding programs [113]. The discovery of the underlying genetic diversity, including its phenotypic-genotypic association, constitutes a challenge for scientific research as well as for society.

1.4.1. DNA markers

Genetic diversity is reflected in DNA markers. Patterns of sequence variations form features that are uniquely linked to a specific locus within the genome. As any genetic diversity, these DNA markers originate from mutations of the nucleotide sequences. They directly reflect the genetic diversity at DNA level and have the benefit of genome-wide distribution. Beside DNA-based molecular markers (DNA-marker) other types of genetic markers exist like morphological markers (phenotypic) or biochemical markers. Morphological markers distinguish an individual or a population from another by phenotypic characteristics like color or shape (e.g. dwarfing) as shown by Kuczyńska et al. [114]. In contrast to that, biochemical markers determine differences at the biochemical level [115] like the tolerance to aluminum in plants [116] or the association of biochemical features with additional subgenome chromosomes in *Aegilops* [117,118]. Both, morphological and biochemical markers have the disadvantages that they occurrence in limited numbers.

In this thesis, I will consider DNA markers in more detail. An example of traditional DNA marker types are restriction fragment length polymorphisms (RFLPs), based on hybridization of DNA-DNA molecules [119]. Besides traditional marker types, several other approaches were established during the last decades [120]. Improvements were accompanied by various automation steps that led to a rapid development of high-throughput genotyping applications [121]. In brief, the most widely used technologies are amplified fragment length polymorphisms (AFLP), simple sequence repeat (SSR), kompetitive allele specific PCR (KASP), diversity array technology (DArT), single nucleotide polymorphism (SNP) and the restriction site associated DNA sequencing (RAD), and genotyping-by-sequencing (GBS) markers. For a general review see [122]. In this brief overview of marker development, the higher automation and throughput has to be emphasized. AFLP development [123] tremendously changed genomic marker development as RAD development did again in 2008 [124]. With the RAD sequencing and genotyping-by-sequencing that was developed in maize and barley [125], restriction enzymes are used to reduce the sequence complexity. Therefore, sequencing is

performed at particular sites assessed by the restriction enzymes. Both approaches combine the process of sequencing and the process of genotyping and therefore are highly cost effective [126]. At present, the sequence-based marker types SNP and SSR are widely used in plant genetic analysis. SNP markers offer the highest throughput and their advantages led to application in many species [127–130]. The power of application in plant research and a detailed description of automated methods, which led into the era of ultrahigh-throughput genotyping are given by Edwards et al. [52].

1.4.2. Single nucleotide variations

The main perspective of this thesis is the study of genetic diversity in plants through the *de novo* detection of single nucleotide variations (SNV). Advances in DNA sequencing technologies have eased and accelerated the discovery of large amounts of SNVs, because sequences data can be generated on low cost for literally any species. The wide dispersion throughout the genome, the high stability and low cost make this DNA marker type very beneficial for application [131]. For variations, that are frequent within a population (abundancy at least 1%) or that are used as markers, the term single nucleotide polymorphisms (SNP) is used [94]. These variations represent genetic diversity at the most fundamental level based on single nucleotide changes. At the DNA level, for a genomic sequence position different sequence alternatives (alleles) are observed. SNVs can be classified into two groups. (1) Transitions are defined as point mutations where a purine is exchanged with a purine or a pyrimidine by another pyrimidine nucleotide. As second type (2) transversions are defined as substitution of a purine by a pyrimidine or vice versa. Another type of variant positions (VP) are INDELs, defined as additional nucleotides (insertions) or lost nucleotides (deletions). For clarity in this thesis, I will use the term variant position as a more general term to refer to both SNV and INDELs. SNVs are the most abundant form of sequence variation within the genome. Consequently, this provides a high density for putative DNA markers in plants, as well as animal or human genomes. Most SNVs occur in non-coding regions and with significant lower frequency in coding regions of genes (exons) [132]. In coding regions, a variation that leads to the change of an amino acid, is defined as non-synonymous SNV. However, through the redundancy in the genetic code a nucleotide change does not always lead to a change in the sequence of the amino acid. A SNV of this type is referred to as synonymous SNV [133]. In non-coding regions a SNV can be classified as intergenic or intronic [134].

With the advance of NGS, the discovery of this type of variation became feasible on a large scale. Furthermore, with the development of high-throughput genotyping platforms, SNVs, and subsequently SNP markers, became one of the predominant used methods during the last years [135–138]. A detailed comparison of SSR and SNP marker technology is given in an empirical study by

Singh et al. [139]. The dense and genome–wide distribution and hence, the possibility to utilize specific markers for any gene of interest, make SNP markers the preferable method for trait analysis. In general, DNA markers will provide access to the genetic diversity within large groups of species. The knowledge about diversity will further support studies of population evolution or association studies. This has already been proven in humans where recent biotechnological achievements paved the way for diversity studies of whole populations [140,141]. These efforts are expected to be applicable to many species. As emphasized by Takeda and Matsuoka [142], the need for crop improvements to overcome the tremendous challenge of nutrition supply requires to transform all this knowledge into plant research. Therefore, they illustrate how concepts like association mapping or quantitative trait loci mapping (QTL), widely used techniques in human research, can support these challenges.

1.5. Computational methods for the detection of diversity

The huge amounts of raw sequence data require a large number of computational calculations to translate pure sequence information into high-quality diversity information. NGS directly paved the way for large scale discovery of SNVs. However, the process to detect a reliable set of valid SNVs is far from being direct and requires several independent sub processes e.g. quality trimming, read alignment, and subsequent variant calling. Each sub process can and should be individually optimized to ensure a reliable discovery of diversity at the end. Variant positions are translated into variant call format (VCF). These files became the standard output format in many variant calling programs. The VCF format goes back to an initiative of the 1000 Genomes Project [143] and its specification was published jointly with a widely used software suite (VCFtools) to manage these files [144]. Various variant calling concepts have been addressed over the last years. These also included the possibility of searching for VPs by cloud-computing [145].

The need for systematic evaluation of variant calling methods has been emphasized in medical applications and was also underlined by a review of published software developments [146]. This publication indicated a low concordance of different variant calling programs and thus motivated a broad evaluation. Figure 4 illustrates the ongoing development of variant discovery tools. The assessment cannot be considered as a complete compendium of available tools, because several tools might still be under development or have not been released in a publication. As a rough estimation, the list of SNV discovery tools provided on SEQanswers [147] contains 80 different tools (October 2014), where a large majority (~85%) are listed as non-commercial. The timeline of software tools (Figure 4) shows the ongoing development of variant calling tools, thus underlining the importance

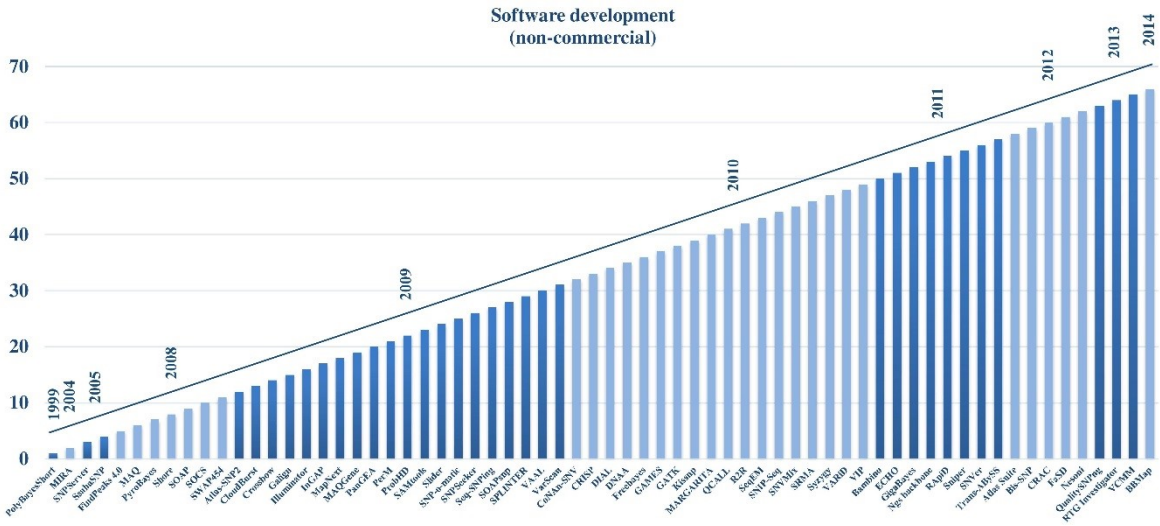

Figure 4 Overview of software tools for variant calling. The graphic is showing the increasing number of non-commercial software tools that are applicable for variant calling. The Illustration uses alternating colors to indicate different years. In consequence two large blocks are emphasized that underline the peak phase of software development (2009 and 2010). In the graphic the continuous increase of available variant calling tools is depicted. In consequence, the high number of different tools is requiring a detail investigation and comparison of predicted results. Motivated by this observation a detailed evaluation for a subset of eight variant calling tools is conducted as part of this thesis.

of variant calling. SNV discovery and its application are discussed in many publications [148,149]. However, these versatile applications further emphasize the need for optimal accuracy and validation. Included in this thesis [2.2] are eight variant callers that demonstrated excellent performance through a broad comparison to other tools [150] or by pre-dominated use within the research community. A survey of diversity detection methods is given by Pabinger et al. [151] as well as a comparison using low-depth sequencing by Xu et al. [152]. However, an extension is required because additional variant calling methods have been released since the study was published. Furthermore, the influence of read alignment to subsequent variant calling was not analyzed in this previous study. As a result, a deeper investigation was required, which is conducted in this thesis.

The most popular SNV calling method currently is SAMtools [153,154]. Several other tools like FaSD [155], VarScan2 [156], VCMM [157], CRISP [158] or SNVer [159] have recently been published and need broad inspection and comparison. In addition, the commercial vendor CLC (http://www.clcbio.com/) is providing a variant calling method that will be analyzed. Finally, the further development of PolyBayes [160] led to a new release and publication of the tool Freebayes [161], which is integrated into this evaluation as well. Referring to Pabinger et al. [151], tools performing genome wide variant discovery are classified in four groups: (i) germline callers, (ii) somatic callers, (iii) tools for the discovery of copy number variation (CNV) and (iv) tools

investigating structural variations (SV). The perspective of this thesis is to reveal constitutional mutations that can be applied as SNP markers. In consequence, the tools reviewed here are classified in the first category. In general, to predict variant sites through reference aligned read files (SAM, BAM or PILEUP), the applied algorithms estimate the probability for a position to be polymorphic. Therefore, Bayesian models are applied in many tools [162–164]. The implementations differ for each developed tool. Nevertheless, for a short explanation the method of VCMM is briefly introduced with reference to an in-depth description in the original publication [157]. For the prediction of a variant site, the probabilistic model calculates a ratio between the probability that the minor allele is an error (P_{error}) and the probability that the minor allele is an alternative allele (P_{allele}). An important influence in this calculation is the base quality (PHRED) score, which reflects the likelihood that a sequencing call is an error. Here, each involved read at the putative variant site is evaluated. When the resulting ratio exceeds the cut-off within the Bayesian decision method, the position is considered to be a variant site. In addition, to improve accuracy of the utilized PHRED scores, another Bayesian decision method can be used beforehand to correct quality scores [165]. Several iterative steps ensure optimized settings for cut-off and quality thresholds. Beside this, several internal analyses check the sequence environment of a putative site, for instance, if the positions are adjacent to INDELs or other variant positions. In fact, each individual variant calling program utilizes independent measurement characteristics and/or slightly different thresholds. As a result, these systematic differences lead to significant differences in variant predictions.

1.5.1. Accuracy of variant discovery

Martin and Wang [166] provide a good review about problems and errors of sequencing technologies that even occur in complexity-reduced RNA-seq data sets. Assemblies are far from being error free and these errors must always be considered in the subsequent diversity study as a particular source of a failure. Measurements to access quality and reliability of variant predictions therefore are of tremendous importance.

Likelihoods of erroneous variant calls are indicated by various tools as scores in the resulting output file. These risk classes provide a good estimation of failure, but comparison of multiple tools often suffers the limitation of different score metrics, particularly when formats differing from the VCF standard are used. Beside this score value, reliability of a prediction can be assessed by several other measurements like: (i) read coverage, (ii) mapping quality, and (iii) minor allele frequency (MAF). In addition to these fundamental measurements, the neighboring quality standard (NQS) was suggested to ensure sufficient quality around a predicted variant position [167]. The NQS criteria

became a widely used measurement and was incorporated in various variant calling tools e.g. GATK [168] or SNPdetector [169].

The complexity of a reliable variant discovery is further increased within highly repetitive plant genomes which are studied in this thesis. In the sequence assembly process reads that originate from repetitive regions often cluster into a single representative contig. Marginal differences within repeats are accepted by the assembler, particularly in WGS assemblies. These collapsed contigs of repetitive sequences can be the source of an error, described as repeat derived false positive calling [170]. Subsequently, variant calling in these repetitive regions will harbor several false positive predictions that are technical artefacts and have no biological meaning. There is a strong requirement, especially for highly repetitive plant genomes, to develop solutions that overcome this issue.

1.5.2. Workflow of diversity filtering

Currently, a massive volumes of NGS data can be used to reveal the genetic diversity of any plant species. These data subsequently generate an unprecedented amount of putative variant sites. Independent of the applied DNA sequencing strategy, the process for variant detection and filtering follows a general workflow, which is depicted in Figure 5. The illustration provides a guideline of existing, extended, and newly developed filter criteria, applied throughout this thesis. It also states in which publication the corresponding criteria have been applied. The presented workflow will serve as a guide through the discussion. It is captured again as retrospective Table 5 with section references and detailed descriptions.

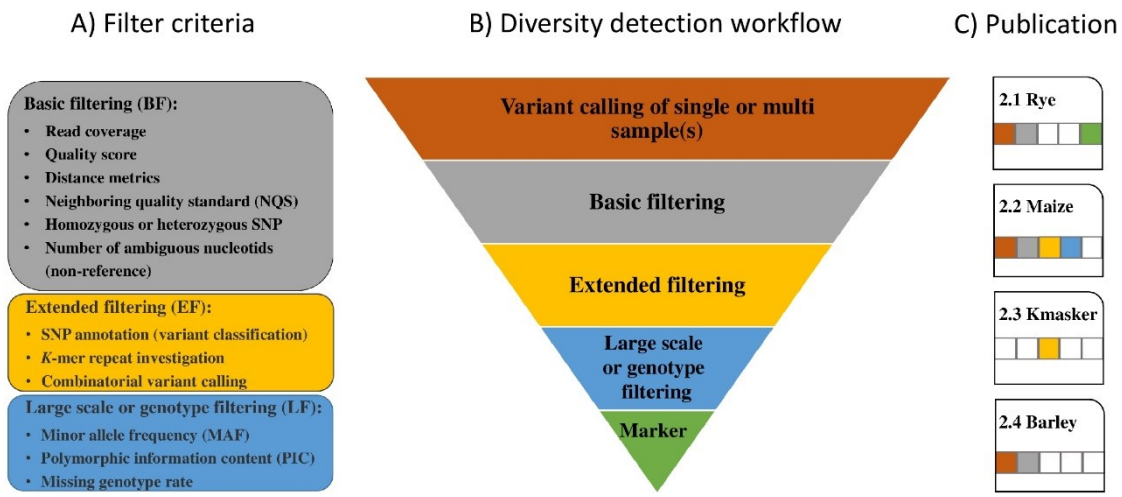

Figure 5 Workflow to filter the descriptive proportion of detected variants. Filtering is embedded subsequent to the process of variant calling (orange) and before the final application of marker development (green). **A)** The listed criteria are classified into the three categories 'basic filtering' (BF), 'extended filtering' (EF) and 'large scale filtering' (LF). The latter one defines criteria applicable to diversity studies with large collection of samples or genotypes. **B)** The filtering methods, depicted in this general workflow, are modular. Different diversity projects might use individual ordering of filter, depending on project aims. **C)** The four publications of this thesis are presented, using the corresponding color code of the diversity workflow, to indicate which methods have been applied.

2. Synopsis of publications

This thesis is based on four independent publications, which together comprise a comprehensive survey of methods for the systematic study of diversity in plants. The main focus is on the economically important crops rye, maize and barley. These crops represent a diverse collection of species with different genome sizes, which range from 2.3 Gbp in *Zea mays*, over 5.1 Gbp in *Hordeum vulgare*, to 7.9 Gbp in *Secale cereale*. In consequence, a reasonably good variety of species is selected to study the complexity and challenges of genome and diversity projects.

2.1. From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.).

Haseneyer*, G., **Schmutzer*, T.**, Seidel, M., Zhou, R., Mascher, M., Schön, C.-C., … Bauer, E. (2011). From RNA-seq to large-scale genotyping - genomics resources for rye (Secale cereale L.). *BMC Plant Biology*, *11*, 131. doi:10.1186/1471-2229-11-131

This article was the first survey of the rye transcriptome leading to a novel sequence and diversity resource. The diversity of rye is studied with an RNA-seq approach analyzing five rye lines by 454 sequencing. I contributed to the project by developing the required bioinformatics pipelines, performing computational analysis, and conducting NGS data processing. Finally the *de novo* assembly, which represents the first transcriptome reference sequence of *Secale cereale*, was conducted by myself. In addition my responsibility was to perform the diversity detection and filtering for high quality SNP markers that were subsequently used for the design of the Rye5k genotyping array. [Contribution 20% and equal contribution (*) of first two authors]

2.2. Targeted Sequencing Reveals Large Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition.

Muraya*, M. M., **Schmutzer*, T.**, Ulpinnis, C., Scholz, U., & Altmann, T. (2015). Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition. *Plos One*, 10(7), e0132120. doi:10.1371/journal.pone.0132120

The project was conceived to study 4,648 biomass-related genes in *Zea mays* using a collection of 21 maize inbred lines. It was my responsibility to design the respective sequence capture array required to perform a CAP-seq experiment. I performed the genomic analysis ranging from *de novo* assembly to variant detection. Furthermore, I comprehensively investigated read alignment and variant calling procedures of multiple programs to conduct an in-depth evaluation of available tools and to establish an enhanced variant calling strategy. The resulting 'combinatorial variant calling' approach increased the overall reliability of the final diversity set. Results indicated a condensed subset of genes supposed to be involved in the biomass constitution in maize. [Contribution 35% and equal contribution (*) of first two authors. Corresponding author].

2.3. Kmasker – a tool for *in silico* prediction of single-copy FISH probes for the large-genome species *Hordeum vulgare*.

Schmutzer*, T., Ma*, L., Pousarebani, N., Bull, F., Stein, N., Houben, A., & Scholz, U. (2014). Kmasker - A Tool for in silico Prediction of Single-Copy FISH Probes for the Large-Genome Species Hordeum vulgare. *Cytogenetic and Genome Research*, *142*(1), 66–78. doi:10.1159/000356460

The problem of repetitive sequences is closely linked with the large and complex genome of barley. My contribution to the publication was the development of an approach to detect unique or low-repetitive sequences that are applicable for fluorescence *in situ* hybridization (FISH). The resulting tool Kmasker is applicable to screen barley sequences for repeat patterns and can be adapted to virtually any species. This precise knowledge of repeats can lead to improved confidence in diversity projects. [Contribution 35% and equal contribution (*) of first two authors]

2.4. A physical, genetical and functional sequence assembly of the barley genome.

IBSC, Mayer, K. F. X., Waugh, R., Brown JW., Schulman, A..,…, **Schmutzer, T.**, …, Stein, N. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, *491*(7426), 711–6. doi:10.1038/nature11543

The publication was conceived by the International Barley Genome Sequencing Consortium (IBSC) and released a comprehensive resource of the barley genome (cultivar 'Morex'), including a physical map of 4.98 Gbp, a majority of which (78.3%) was anchored to a high-resolution genetic map. In addition, an extensive genome-wide study of the natural diversity in domesticated and wild barley was performed. I shared the responsibility for the manuscript subsection 're-sequencing and diversity analysis'. With respect to that, I performed the genomic data processing of the four re-sequenced barley cultivars 'Bowman', 'Barke', 'Igri', and 'Haruna Nijo', as well as the progenitor line *Hordeum vulgare ssp. spontaneum*, revealing the preliminary source of variant positions. [Contribution 2%]

3. Publications

Haseneyer et al. BMC Plant Biology 2011, 11:131
http://www.biomedcentral.com/1471-2229/11/131

BMC Plant Biology

RESEARCH ARTICLE Open Access

From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.)

Grit Haseneyer[1†], Thomas Schmutzer[2†], Michael Seidel[3], Ruonan Zhou[4], Martin Mascher[2], Chris-Carolin Schön[1], Stefan Taudien[5], Uwe Scholz[2], Nils Stein[4], Klaus FX Mayer[3] and Eva Bauer[1*]

Abstract

Background: The improvement of agricultural crops with regard to yield, resistance and environmental adaptation is a perpetual challenge for both breeding and research. Exploration of the genetic potential and implementation of genome-based breeding strategies for efficient rye (*Secale cereale* L.) cultivar improvement have been hampered by the lack of genome sequence information. To overcome this limitation we sequenced the transcriptomes of five winter rye inbred lines using Roche/454 GS FLX technology.

Results: More than 2.5 million reads were assembled into 115,400 contigs representing a comprehensive rye expressed sequence tag (EST) resource. From sequence comparisons 5,234 single nucleotide polymorphisms (SNPs) were identified to develop the Rye5K high-throughput SNP genotyping array. Performance of the Rye5K SNP array was investigated by genotyping 59 rye inbred lines including the five lines used for sequencing, and five barley, three wheat, and two triticale accessions. A balanced distribution of allele frequencies ranging from 0.1 to 0.9 was observed. Residual heterozygosity of the rye inbred lines varied from 4.0 to 20.4% with higher average heterozygosity in the pollen compared to the seed parent pool.

Conclusions: The established sequence and molecular marker resources will improve and promote genetic and genomic research as well as genome-based breeding in rye.

Keywords: EST resource, next generation sequencing, *Secale cereale* L., Rye5K SNP array, single nucleotide polymorphisms

Background

The improvement of agricultural crops with regard to yield, resistance and environmental adaptation is a perpetual challenge for both breeding and research. With regard to prospected climate changes, improved tolerance against abiotic stresses like drought, low soil fertility, and extreme temperatures is required in crop improvement. The outcrossing species rye shows the highest freezing tolerance among small grain cereals [1] and exhibits excellent tolerance against many biotic and abiotic stresses. Understanding the functional genetic basis of stress tolerance in rye will facilitate the improvement of stress tolerance in wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.). As a genetic research system, rye is intriguing due to its high genetic variability. In addition to being an economically important crop for Middle and Eastern Europe, rye provides valuable traits for other crops, as a parent of the amphiploid triticale, and as a donor of translocated chromosome segments in wheat [2]. Rye benefits from being diploid and closely related to the more extensively characterized species wheat and barley. While reference sequences of grass genomes have become available for rice [3,4], sorghum [5], *Brachypodium* [6] and maize [7], sequence information for rye is sparse which hampers the exploitation of its genetic potential.

The haploid genome size of rye is more than 8 Gbp [8] which is one of the largest among cereal crops. In addition, 92% of the genome is composed of repetitive sequences [9]. Genetic and genomic resources are limited compared to other *Triticeae*. Currently, 1,073,668 wheat and 501,620 barley ESTs are publicly available

* Correspondence: eva.bauer@wzw.tum.de
† Contributed equally
[1]Plant Breeding, Technische Universität München, Centre of Life and Food Sciences Weihenstephan, 85354 Freising, Germany
Full list of author information is available at the end of the article

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

whereas only 9,298 rye ESTs are deposited in public databases http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html (release 070111). Publicly available genomic resources for rye are restricted to one BAC library [10], a limited number of genetic markers http://wheat.pw.usda.gov/GG2/index.shtml, and genetic maps with low marker density [11-15].

Next-generation sequencing (NGS) technologies such as Illumina's Genome Analyzer and Roche's 454 sequencing platforms have opened the way to tackle sequencing of large genomes like those of barley and wheat which would be impossible to address by Sanger sequencing [16]. NGS platforms produce hundreds of thousands of sequences in a massively parallel manner, are cost and labour effective and were proven to be reliable and accurate. Several studies have highlighted the success and usefulness of NGS for extending available genomics resources by transcriptome [e.g. [17,18]] and whole-genome [19] sequencing. Furthermore, NGS has been used for gene expression profiling [20], analysis of genome organisation [21], DNA methylation studies [22], and molecular marker development [23], to name few.

Given the large genome size and the lack of sequence information and genomic resources in rye, identification and targeted isolation of genes underlying agronomic traits and understanding of gene function and trait variation is greatly hampered. The aim of the present study was to promote rye genome analysis through massive improvement of the public rye EST resource and development of the first high-throughput SNP genotyping array.

Methods

Plant material, RNA and sequencing

Five winter rye inbred lines Lo7, Lo152, Lo225, P87, and P105 were used for cDNA sequencing. Lo7, Lo152, and Lo225 were provided by KWS LOCHOW GMBH (Bergen, Germany) and represent lines from the seed parent and the pollen parent pool of the company's hybrid rye breeding program. P87 and P105 were developed at the Institute of Genetics and Cytology, Minsk, Belarus, and are parents of the mapping population P87 × P105 [24]. Inbred lines Lo7, Lo152, and Lo225 were generated by six selfing generations, whereas P87 and P105 were selfed seven and eight times, respectively. In addition, 54 proprietary inbred lines from the breeding material of KWS LOCHOW GMBH, representing the two breeding pools were investigated. Lines from the pollen parent pool were generated by two to three selfing generations, whereas lines from the seed parent pool have undergone five selfing steps.

To capture a comprehensive part of the rye transcriptome 20 samples of total RNA per inbred line were obtained from a set of plant tissues harvested at five developmental stages and after three stress treatments, respectively (Additional file 1). Three plants per inbred line were pooled to obtain each of the 20 RNA samples. For all non-stress treatments tissue samples from leaves, stems and/or roots were harvested at three- to four-leaf stage, tillering, stem extension, heading and harvest ripe stage. Coleoptiles, florets, early and mature spikes were harvested. To enrich stress induced genes in the cDNA sample, cold stress, dehydration shock, and nutrient-starvation stress treatments were applied in the three- to four-leaf stage. Cold stress was induced by placing plants in a freezer at -15°C. Root, stem and leaf tissues were harvested after 1, 3, and 6 hours of stress treatment and pooled. Dehydration shock experiments were conducted by removing well-watered plants from soil and leaving them on Whatman® 3 MM paper (Whatman GmbH, Dassel, Germany) at room temperature [25]. Root, stem, and leaf tissues were harvested after 3, 6, and 12 hours of stress and pooled. Three plants per inbred line were densely planted leading to nutrient-starvation stress. Root and leaf tissues were harvested and pooled. All tissue samples were frozen in liquid nitrogen and stored at -80°C until use. Total RNA was isolated according to manufacturer's instructions using the NucleoSpin RNA Plant kit (#740949, Macherey-Nagel, Düren, Germany) and quantified with the SPECTRONIC GENESYS™ 10 BIO spectrometer (Thermo ELECTRON CORPORATION, Madison, USA).

Five micrograms of the 20 RNA samples of each inbred line were pooled and 100 μg total RNA per inbred line was sent for cDNA synthesis to vertis Biotechnology AG (Freising, Germany). Poly(A)+ RNA was prepared from total RNA. First-strand cDNA synthesis was primed with random hexanucleotide primers. Then 454 sequencing adapters A (5'-GCCTCCCTCGC GCCATCAG-3') and B (5'-CTGAGCGGGCTGGCA AGGC-3') were ligated to the 5' and 3' cDNA ends. Finally, cDNAs were amplified in 20 (Lo152) and 21 (Lo7, Lo225, P87, P105) PCR cycles using a proof reading enzyme. Normalization was carried out by one cycle of denaturation and reassociation of the cDNA. Reassociated ds-cDNA was separated from the ss-cDNA on hydroxylapatite columns to obtain the normalized cDNA samples. After hydroxylapatite chromatography, the ss-cDNA samples were amplified in 8 PCR cycles. The cDNA fraction in the size range of 600 to 800 bp was eluted from preparative agarose gels. As a control, aliquots of the fractionated cDNAs were analyzed on 1.5% agarose gels. Approximately 150 to 250 μg of the normalized, adapter-ligated, and size selected cDNA samples were used for GS FLX 454 sequencing. All 454 sequence raw data were submitted to the EBI sequence

read archive (SRA) and are available under the study accession number ERP000274.

EST resource

De novo sequence assembly

After 454 sequencing, raw sequence reads were passed through quality filtering where cDNA synthesis primer and sequencing adapter sequences were removed. After pre-processing, cleaned and trimmed reads were subjected to inbred line-specific assemblies. Therefore, we adapted the strategy of Kumar and Blaxter [26] for assembling transcriptome data using multiple assembly programs and combining the outcomes to create longer contigs that are less likely to be *in-silico* artefacts brought forth by a single algorithm. The strategy has been modified to be applicable for various lines (Figure 1). We used three independent assemblers to achieve most credible consensus contig sequences. Initially, all reads from each of the five lines were assembled separately into first-order contigs with the programs CLC assembly cell v3.20 http://www.clcbio.com, Mira v3.21 [27] and Newbler v2.5 [28]. While MIRA and Newbler follow the overlap-consensus-layout paradigm (OLC), CLC attempts to find paths in De Bruijn graphs. To obtain line-specific assemblies, all first-order contigs constructed by the three assemblers were merged using the OLC assembler CAP3 [29]. We considered only line-specific contigs whose constituents included first-order contigs from all three assemblers. For EST resource generation (Sce_Assembly03), we employed CAP3 a second time to co-assemble the high confidence line-specific contigs and denoted those supported by constituents from more than one line as multi-line contigs, while contigs with evidence from only one line were deemed single-line contigs. The assembly process of Sce_Assembly03 has been accomplished with a screening for potential DNA and foreign RNA contamination. We applied a BlastN against chloroplast genome sequences of barley (GenBank: NC_008590) and wheat (GenBank: NC_002762), mitochondrial genome sequences of rice (GenBank: AP011077), sorghum (GenBank: DQ984518), and wheat (GenBank: GU985444), and plastids genome sequences of *Brachypodium* (GenBank:EU325680), rice (GenBank: GU592207), sorghum (GenBank: NC_008602), and wheat (GenBank: AB042240). Further purity was gained by excluding hits against CDS sequences of *Acyrthosiphon pisum* (GenBank: ACFK00000000), *Buchnera aphidicola* (GenBank: AE013218), *Fusarium graminearum* (GenBank: AACM00000000), and the draft sequence of *Puccinia triticina* available at the Broad Institute. We discarded contigs from the Sce_Assembly03 sequence set that showed E-values larger than E-20 and the proposed best hits representing at least 10% of the full contig size. The established EST resource Sce_Assembly03 is available from the GABI primary database [30], http://www.gabipd.org.

Sequence comparisons

Sequences between the five rye inbred lines potentially differ to a degree that prevents the *de novo* assembly of two lines. Blast [31] comparisons which do not require strict sequence identity were carried out to analyze for overlaps between the different assemblies. Line-specific assemblies generated by CAP3 were used together with the Sce_Assembly03 in an "all versus all" BlastN analysis. Each line-specific assembly as well as the multi-line and single-line contigs of the Sce_Assembly03 were used as both, subject and query sequences. The best query hit to a subject sequence was counted to identify homologs in the respective assemblies. Hits were considered significant when they exceeded a conservative cut-off value of > = 70% identity and 30 bp coverage.

Comparisons of the Sce_Assembly03 against the four currently available protein databases of maize [ZmB73_v5b.60, http://www.maizesequence.org], rice [RAP2, [32]], sorghum [5], and *Brachypodium* [6], two EST databases of barley and wheat (Barley assembly 35 and Wheat assembly WK, http://harvest.ucr.edu), and two full length cDNA (flcDNA) library databases of barley [33] and wheat [34] were performed using BlastX and tBlastX, respectively. Hits were only considered significant when they exceeded a conservative cut-off value of > 70% identity and 30 bp coverage. To prevent hits found based on low-complexity sequences or repeats the Sce_Assembly03 was masked using RepeatMasker [35] and the internal MIPS repeat database [36].

Genome-wide distribution of the Sce_Assembly03 contig sequences was investigated by chromosome-wise BlastX analysis comparing multi-line and single-line contigs with *Brachypodium* protein sequences. Sce_Assembly03 sequences were mapped onto the *Brachypodium* genome by using a sliding window approach with a window size of 0.5 Mb and a shift of 0.1 Mb along the *Brachypodium* chromosomes. The number of BlastX hits and the percent bp coverage of the respective *Brachypodium* genes were determined. These density values were corrected for the number of Ns per window, if the N content exceeded 60% the value was set to zero. Density values were extrapolated to genes [6] or hits (rye) per Mb to facilitate comparisons. To visualize the mapping results heatmaps were created from the density values using the Python matplotlib module in combination with the jet colormap [37].

Functional gene annotation

The 115,400 sequences of the Sce_Assembly03 were functionally annotated performing a Blast search with Blast2GO default parameters against the non-redundant (nr) protein sequence database [38] after masking

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

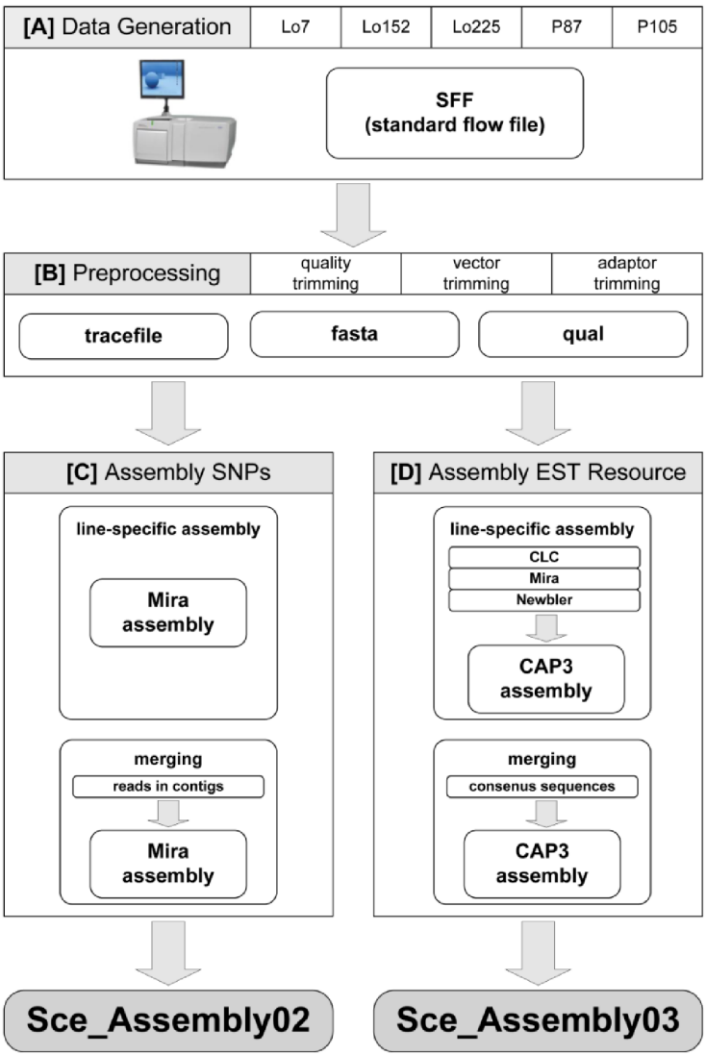

Figure 1 Pipeline for the assembly procedure of Roche/454 sequence reads. After data generation [A], sequence (fasta), quality (qual) and trace file information were extracted. Low quality regions, vector and adaptor sequences were removed from raw reads [B]. Preprocessing was finished by subjecting trimmed reads to the line-specific assembly. For establishment of the SNP resource Sce_Assembly02 [C] only reads assembled in contigs of line-specific assemblies were subjected to the merging process of the second assembly using Mira. For establishment of the EST resource Sce_Assembly03 [D] assemblies were computed for each of the five lines separately with CLC assembly cell, Mira, and Newbler and merged by CAP3 assembly. Consensus sequences of all lines were passed to a second CAP3 assembly combining sequences over multiple lines. The resulting sequence set comprises contigs that were confirmed by consensus sequences from two to five lines (multi-line contigs) or contigs that contain reads originating from one line (single-line contigs).

repetitive sequences and excluding the singletons. Gene ontology (GO) terms were assigned using B2G4PIPE http://www.blast2go.org and a locally installed Blast2GO database. The annotation file was extended by its respective GO category - biological process, cellular component, and molecular function - using a custom built Python script that is available upon request.

SSR mining and SNP discovery

Simple sequence repeat (SSR) motifs within 338,536 contigs of the line-specific assemblies were identified by

MISA [39] under standard settings. Out of the five inbred lines, Lo225 was selected as reference dataset as it provided the highest number of SSR containing contigs. The MISA output of the four remaining lines was cross-matched with the Lo225 dataset to detect redundant SSRs. A non-redundant SSR dataset was generated by combining "unique" SSR motifs detected in Lo7, Lo152, Lo225, P87, and P105. Mononucleotide repeat motifs were discarded since monomer runs are known to be the most frequent sequencing errors in Roche/454 data. For experimental validation of *in silico* detected SSRs, primers flanking the SSR motifs were designed using Primer3 [40]. Amplification of the fragments was performed in Lo7, Lo225, P87, and P105 as they are the parents of two mapping populations. Thus, polymorphisms detected between Lo7 and Lo225 and/or P87 and P105 enable the genetic mapping of discovered SSRs. PCR was conducted in a total volume of 20 μl, including 20 ng of genomic DNA, 1× HotStar Taq PCR buffer (Qiagen, Hilden, Germany), 250 nM of each primer, 200 μM dNTPs, and 0.5 U HotStar Taq DNA polymerase (Qiagen, Hilden, Germany). Using a touch-down PCR profile, an initial denaturation step of 15 min at 95°C was followed by 45 cycles of denaturation at 94°C for 1 min, annealing for 1 min, and extension at 72°C for 1 min. Annealing temperature was decreased by 1°C per cycle from 65°C to 55°C and was kept constant for 35 subsequent cycles. A final extension step was performed at 72°C for 10 min. Successful amplification was checked on 1.5% agarose gels.

For the discovery of SNPs in assembled sequences, a second assembly strategy was pursued. Reads assembled in line-specific contigs were selected from all reads and subjected to an overall assembly, merging the extracted reads of all five genotypes (Sce_Assembly02, Figure 1). With this strategy information about nucleotide coverage is maintained which is important for reliable SNP discovery. The Sce_Assembly02 is described in Additional file 2 and is available from the GABI primary database http://www.gabipd.org. The workflow from *in silico* SNP discovery in the Sce_Assembly02 to selection of high confidence SNP candidates was a three-step procedure: First, the tool GigaBayes V0.4.1 [41] was applied with parameter settings given in Additional file 3. Second, characteristics for discovered SNPs were extracted by in-house implementations to compute defined selection criteria for candidate SNPs. Candidate SNPs were filtered by these selection criteria to meet the following requirements: SNPs should be bi-allelic and polymorphic between parents of the two mapping populations Lo7 × Lo225 and/or P87 × P105. For successful probe design they should have a distance to homopolymeres > 5 bp, to the next Indel > 60 bp, and to the contig end > 60 bp. Third, filtered SNPs were manually inspected in the assembled sequences using EagleView [42] to ensure high quality of the SNP genotyping array. We considered putative sequencing errors, SNP position in individual reads, and haplotype information. Oligoprobes for 5,234 SNP were designed and the Rye5K array was produced by Illumina Inc. (San Diego, USA) as Infinium iSelect HD Custom BeadChip. To demonstrate genome-wide coverage of the SNPs represented on the genotyping array SNP containing contig sequences were *in silico* mapped against the *Brachypodium* genome by BlastN analysis.

SNP array performance was assessed by analyzing 59 rye inbred lines including the five inbred lines used for sequencing as well as accessions from barley (Barke, Morex, OWB Dom, OWB Rec, Steptoe), wheat (Chinese Spring, Dream, Mulgara), and triticale (Modus, breeding line SaKa3006). A total of 300 ng genomic DNA per plant was used for genotyping on the Illumina iScan platform and the Infinium HD assay following manufacturer's protocol. The fluorescence images of an array matrix carrying Cy3- and Cy5-labeled beads were generated with the two-channel scanner. Raw hybridization intensity data processing, clustering and genotype calling (AA, AB, BB) were performed using the genotyping module in the GenomeStudio software V2009.1 (Illumina, San Diego, USA). Genotype data were cleaned through exclusion of all SNP assays with more than 5% missing data. Frequencies of the A and B allele for a given SNP were calculated directly by dividing the number of occurrences of one allele (AA + 1/2 AB or BB + 1/2 AB) by twice the number of assayed lines per SNP. Residual heterozygosity of 59 inbred lines was calculated by the relation of heterozygous SNPs (AB) to the number of assayed SNPs per inbred line. Significant deviation of the observed value from the expected value was tested with an exact binomial test using R [43]. Genotyping data of the 10 non-rye accessions were analyzed to investigate the applicability of the Rye5K SNP array to other small grain cereals.

Results

Establishment and description of the rye EST resource

Assembly

The five independent sequencing runs produced between 364,343 and 681,787 reads corresponding to ~87 and ~166 Mb of raw data per inbred line (Table 1). Subsequent quality filtering and removal of sequencing adapters and cDNA synthesis primers resulted in ~75 to ~145 Mb of high quality sequences per inbred line with median read lengths between 213 and 222 bp. Overall, 2,573,590 high quality reads with a median length of 216 nucleotides were obtained, totalling 548 Mb. The quality filtered reads of the five line-specific cDNA libraries were assembled separately generating between

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

Table 1 Descriptive statistics of five independent Roche/454 GS FLX sequencing runs

	Inbred line				
	Lo7	**Lo152**	**Lo225**	**P87**	**P105**
Raw sequence data					
Number of sequences	364,343	469,345	572,518	488,829	681,787
Average read length [bp]	239	248	242	240	244
After quality filtering					
Number of sequences	363,681	469,208	571,433	488,132	681,136
Average read length [bp]	207	220	213	208	214
Total bp	75,281,967	103,225,760	121,715,229	101,531,456	145,763,104
25% quantile [bp]	203	210	208	203	207
Median [bp]	213	222	218	213	217
75% quantile [bp]	223	236	229	223	228

51,462 and 78,813 contig sequences per line-specific assembly, summing up to 338,536 contigs (Additional file 2). On average each nucleotide in the five line-specific assemblies was covered by 4.5 to 6.2 reads.

Consensus sequences created by multiple assembly programs and merged by CAP3 were used to generate the Sce_Assembly03 (Figure 1, Table 2). 89.0% of the reads were assembled into contigs originating from two, three, four, or five inbred lines (multi-line contigs) or from one single inbred line (single-line contigs), respectively. The Sce_Assembly03 resulted in 115,400 sequences including 33,352 multi-line contigs (77.8% of all reads) and 82,048 single-line contigs (11.1% of all reads). 11.0% of all reads failed the quality criteria and were removed from the assembly. The multi-line contig sequence length ranged from 201 bp to 8,636 bp with a L50 length of 1,070 bp. On average, each contig was built from sixty reads in the multi-line contigs and three reads in the single-line contigs.

Sequence comparisons

We compared the five line-specific assemblies generated by CAP3 against each other and against the multi-line and single-line consensus sequences of the Sce_Assembly03

Table 2 Description of the Sce_Assembly03

	Multi-line contigs	**Single-line contigs**
Number of reads	2,000,855	286,386
Number of reads/contig	60	3
L30 [bp]	1,527	505
L50 [bp]	1,070	333
L70 [bp]	727	247
Number of contigs	33,352	82,048
< 500 bp	11,188	71,581
501-1000 bp	12,679	8,347
1001-2000 bp	7,693	1,952
2001-5000 bp	1,767	166
> 5000 bp	25	2
Longest sequence [bp]	8,636	5,721

(Table 3). This revealed 52.16% to 78.72% hits between the line-specific assemblies. BlastN analysis of the line-specific assemblies against the multi-line contigs reached up to 87.79% hits. Thus, as expected, a large overlap of represented genes between single-line assemblies can be concluded. However, the remaining 12.21% revealed either pronounced sequence differences (highly polymorphic genes/alleles) or genes that are represented (expressed) in only one of the five rye inbred line samples.

The sequence homology between the line-specific assemblies and the Sce_Assembly03 with the reference genomes of *Brachypodium*, maize, rice, and sorghum, and available flcDNA and EST collections from wheat and barley, respectively, was investigated by (t)BlastX comparisons (Figure 2). Most homologs were identified in comparison to barley sequences, followed by *Brachypodium*, wheat, sorghum, maize and rice. Contig sequences of the line-specific assemblies and multi-line contigs of the Sce_Assembly03 showed a high homology to the public sequence databases. Low homology was detected for the single-line contigs of the Sce_Assembly03. This finding can be attributed to the sequence length which is about two thirds shorter than that of multi-line contigs (Table 2). Multi-line contigs of the Sce_Assembly03 yielded more than 65% hits with either barley or wheat flcDNA and HarvEST assemblies (data not shown). Through tBlastX comparisons of the Sce_Assembly03 against the genome sequences of *Brachypodium*, maize, sorghum, and rice we were able to tag fragments from about 46.3%, 35.9%, 37.2% and 36.2% of the reference gene repertoires. From 33,352 multi-line and 82,048 single-line contigs 22,926 (68.7%) and 23,406 (28.5%) revealed a hit to at least one of the public grass sequence resources. The genes comprised in the rye cDNA libraries indicated no bias for or against a certain region of the rye genome when comparing the Sce_Assembly03 contig sequences to the *Brachypodium* genome (Additional file 4). The dense gene content in the distal regions of the *Brachypodium*

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

Table 3 BlastN comparisons of the five line-specific assemblies generated with CAP3 and the Sce_Assembly03

	Query						
	Line-specific assembly					Sce_Assembly03	
Subject	Lo7	Lo152	Lo225	P87	P105	Multi-line contigs	Single-line contigs
Line-specific assembly							
Lo7		52.2	56.1	61.8	56.9	76.1	35.5
Lo152	67.7		54.3	59.6	56.0	77.1	49.5
Lo225	77.6	58.3		68.7	63.8	84.2	53.5
P87	74.4	55.4	59.9		60.9	82.8	40.6
P105	78.7	59.5	63.8	70.2		87.8	47.5
Sce_Assembly03							
Multi-line contigs	85.2	64.4	69.6	78.0	72.3		35.3
Single-line contigs	59.1	64.4	67.3	59.2	62.4	58.5	

Values show percent hits of query sequences counting the first best hit in each comparison.

chromosomes as well as the gene poor regions around the centromeres were well covered by Sce_Assembly03 contig sequences.

Functional gene annotation

After masking repetitive sequences of the Sce_Assembly03 111,150 sequences (32,725 multi-line and 78,425 single-line contigs) remained for Blast2GO analysis. Out of these sequences 49,294 revealed a hit against the nr database and subsequently 35,356 (71.7%) unique rye contig sequences (16,970 multi-line and 18,386 single-line contigs) were assigned to one or more GO annotations. In total 35,186, 38,280 and 51,950 GO terms were obtained for biological processes, cellular components and molecular functions, respectively (Additional file 5).

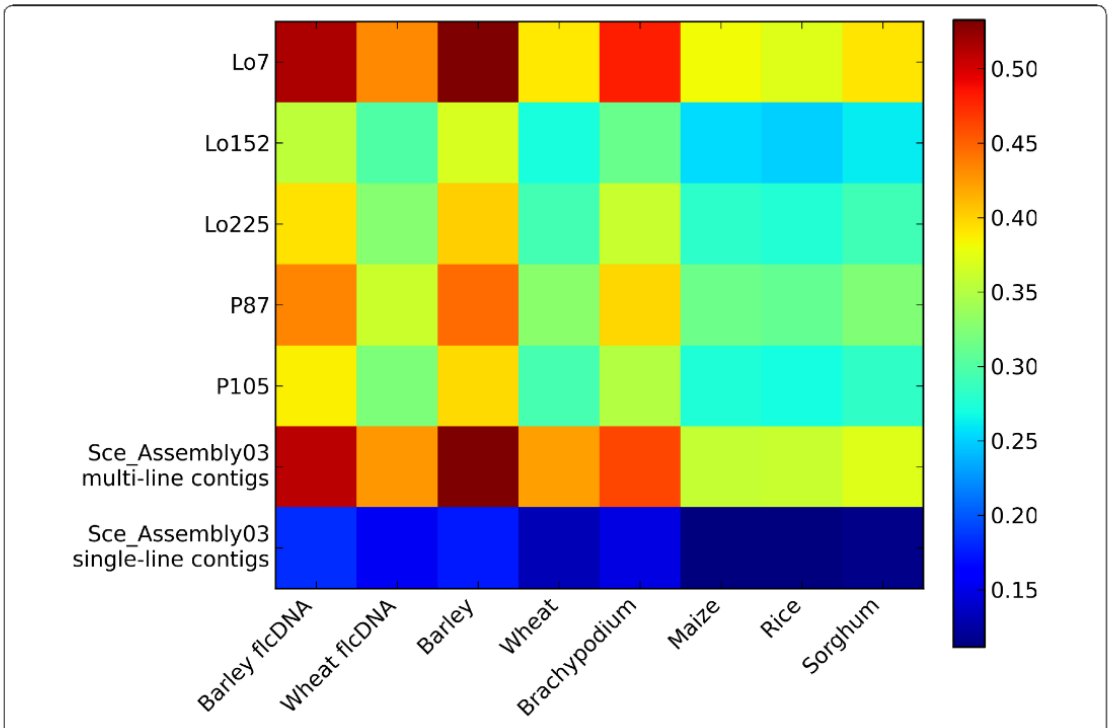

Figure 2 Heatmap of (t)BlastX analysis results to public model grass genomes and *Triticeae* EST and full length cDNA (flcDNA) resources. Contig sequences from the line-specific assemblies generated by CAP3 and the Sce_Assembly03 were aligned to public barley and wheat EST and flcDNA sequences and to *Brachypodium*, maize, rice, and sorghum genomic sequences. Percent hits to individual databases were counted using a 70% similarity cutoff and visualized in colours (colour code shown on the right).

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

Across the three GO categories, 4,997 unique GO terms were identified. More than 350 sequences in the Sce_Assembly03 were related to biotic and abiotic stress response (data not shown).

Marker discovery, SNP array design and high-throughput genotyping

SSR marker development

Within the 338,536 contigs of the line-specific assemblies a fraction of 12,317 (3.6%) contigs contained SSR motifs. Primer sequences could be designed for 5,230 of these contigs. Restriction to di-, tri-, tetra-, penta- or hexa-nucleotide motifs reduced the number of SSR candidates to 3,799. Cross-match analysis filtered a final SSR dataset comprising 1,385 unique, non-redundant SSRs (Additional file 6). A random subset of 155 SSRs was chosen for experimental validation by PCR amplification of the four parental genotypes Lo7, Lo225, P87, and P105. 146 primer pairs (94%) immediately amplified fragments of expected size without further optimization of PCR conditions. Twelve primer combinations produced fragments larger than expected indicating the presence of introns. These were excluded from further analyses. Finally, 61 (46%) out of 134 PCR products with expected fragment size revealed naked-eye polymorphisms on agarose gels between either P87 and P105 (29) or Lo7 and Lo225 (37).

SNP discovery

SNP discovery requires sufficient coverage with high quality sequence reads in order to allow for distinguishing true SNPs from sequencing errors. Therefore, the assembly Sce_Assembly02 was performed that excluded singletons from the line-specific assemblies when merging sequences of the five inbred lines. Overall 277,033 putative polymorphisms in 138,339 contigs cumulating 55 Mb consensus sequences were identified in a first data mining step using GigaBayes. The number of SNP candidates was reduced to 17,917 by filtering those SNPs that fulfilled the selection criteria and quality requirements such as bi-allelic and polymorphic between parents of the two mapping populations Lo7 × Lo225 and/or P87 × P105, distance to homopolymeres > 5 bp, distance to the next Indel > 60 bp, and distance to the contig end > 60 bp. Subsequent manual inspection in the Sce_Assembly02 reduced the dataset to 5,211 SNP candidates from 3,961 contigs. This dataset together with additional 23 SNPs discovered in non-public rye sequences was used for the design and production of the Rye5K SNP genotyping array. Out of the 3,961 unique contigs, 2,835 contigs (71.6%) were *in silico* mapped to the *Brachypodium* genome. The contigs were evenly distributed with 826, 641, 688, 416, and 262 hits on chromosomes Bd1 to 5, respectively (Additional file 4). Blast2GO analysis of 3,961 contig sequences represented on the Rye5K array assigned 2,096 sequences with associated GO identifiers (Additional file 7).

Application of the Rye5K SNP array

The performance of the Rye5K SNP array was tested on the five inbred lines selected for RNA-*seq*, 54 additional rye inbred lines, and 10 non-rye accessions. Out of the 5,234 SNPs, 4,557 (87%) generated signals and between 2,970 (57%) and 3,148 (60%) were successfully called for the 59 rye inbred lines representing the hybrid rye seed parent and pollen parent pools (Table 4 Additional file 8). Based on genotyping results for the five inbred lines used for SNP discovery, 3% of the *in silico* detected SNPs turned out to be false positives. Allele frequencies in rye were evenly distributed from 0.1 to 0.9 (Figure 3). A small proportion of 12.3% called SNPs turned out to be monomorphic in the independent set of 54 inbred lines not used for SNP discovery with slightly increasing values when looking separately at the pollen parent (15.7%) and the seed parent (13.7%) pools.

Genotyping data were used to calculate the observed residual heterozygosity of the rye inbred lines. The observed percentage of heterozygous loci for each line varied between 4.1 and 4.8% in the five rye inbred lines used for 454 sequencing and between 4.0 to 20.4% in the 54 inbred lines from the two heterotic breeding pools. On average, a higher level of residual heterozygosity was observed for the pollen parent pool (11.5%) than for the seed parent pool (5.5%).

Applicability of the Rye5K SNP array to other small grain cereals was investigated. Out of the 4,557 SNP assays that generated a signal in rye, 63.0% (2,871), 75.8% (3,452), and 84.1% (3,831) could be scored in barley, wheat, and triticale, respectively. However, 86.7, 91.6, and 76.5% of the scored SNPs did not show a polymorphism between the investigated barley, wheat, and triticale accessions.

Discussion

Dual-purpose transcriptome sequencing

In this study we report the establishment of rye genomic resources comprising 115,400 EST sequences, 1,385

Table 4 Heterozygosity of five sequenced rye inbred lines after genotyping with the Rye5K array

	Inbred line				
	Lo7	Lo152	Lo225	P87	P105
Loci total	3,145	3,133	3,134	3,148	3,127
Homozygous loci	3,004	3,005	2,987	2,997	2,988
Heterozygous loci	141	128	147	151	139
Generation	F_7	F_7	F_7	$F_{7:10}$	$F_{6:9}$
Expected heterozygosity [%]	1.6	1.6	1.6	1.6	3.1
Observed heterozygosity [%]	4.5***	4.1***	4.7***	4.8***	4.4*

Significant (***: *p*-value < 0.01, *: *p*-value < 0.05) deviation from the expected level of heterozygosity is indicated.

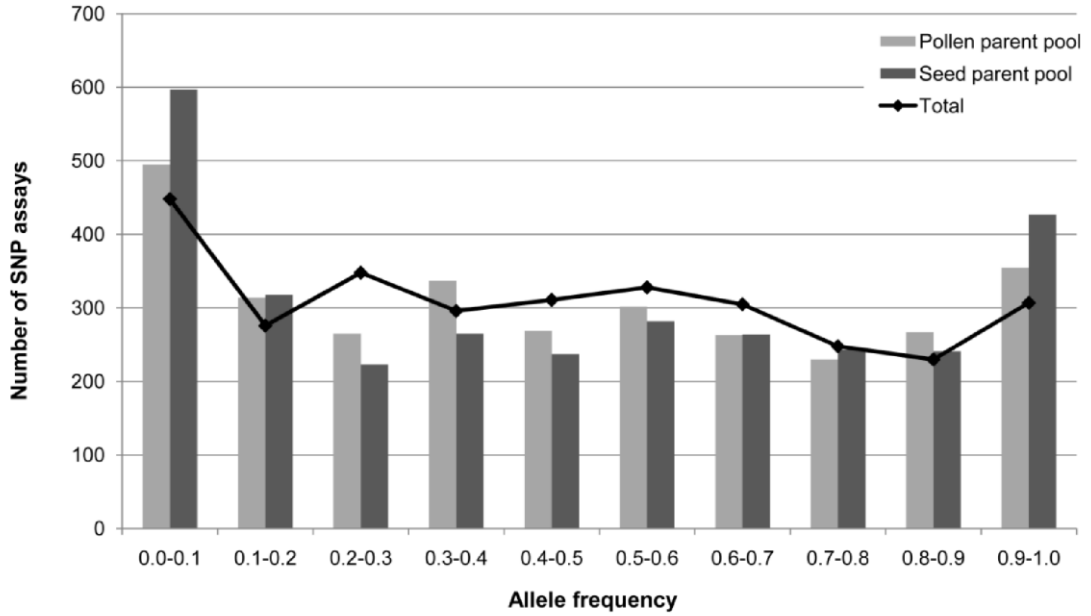

Figure 3 Distribution of allele frequencies for evaluable SNPs on the Rye5K SNP array. Allele frequencies observed in total and separately in the rye breeding seed parent and pollen parent pools belong to one category if the value is > the left category border and ≤ the right category border. Allele frequency values equal to 0 and 1 fall into the first and last category, respectively.

SSRs, more than 5,000 SNPs, and the Rye5K SNP array for large-scale genotyping. NGS was used to generate transcriptome sequences of the five rye inbred lines Lo7, Lo152, Lo225, P87, and P105. The number of reads per sequencing run of the present study was in line or even surpassed results obtained in other studies [17,23,44]. Due to the massive number of 2.5 Mio read sequences obtained by 454 sequencing the *de novo* assembly of such datasets remains a computational and bioinformatic challenge. Two purpose-oriented assembly strategies were followed in order to first provide a comprehensive EST resource and second enable discovery of polymorphisms between inbred lines. A second assembly on top of the five line-specific assemblies reduced the possibility of creating chimeric artefacts in the Sce_Assembly03. In addition, sequence redundancy introduced by variations between lines is removed. This was achieved by bringing together related sequences while accepting line specific nucleotide differences. In contrast this fact was essential for SNP detection, where only reads that were pre-assembled in line-specific contigs were subjected to the Sce_Assembly02. Thus, information about allele coverage at the SNP position was retained which increased the reliability of SNP candidates. A challenge in our study was the detection of SNPs without a reference sequence. Many SNP detection tools such as GMAP [45] or MAQ [46] are only applicable to *de novo* assemblies that are aligned to a reference sequence. This was a strong challenge in our approach and much effort was invested in the detection of high confidence SNPs. Manual inspection of SNP candidates in more than 10,000 contigs indicated that many sequencing errors occurred in the beginning of read sequences which, as a consequence, lead to false positives. Exclusion of SNP candidates detected in such regions of read sequences might reduce the false positive rate and improve automated tools that detect polymorphisms in *de novo* assembled sequence data without a reference sequence.

Genome sequencing has progressed rapidly in model plants. Given the increased sequencing throughput and the decreasing costs, NGS technologies pave the way for sequencing even large genomes [47-49]. Although of major importance for research and breeding, sequence resources for rye were sparse imposing serious limitations for trait mapping, association studies, and functional genomics in rye. Rye is of interest especially for Middle and Eastern European economic markets due to its high tolerance to abiotic stresses. As a first step towards deciphering the rye genome we aimed to sequence a large portion of the rye transcriptome. To achieve this we first sampled RNA from plants under

Haseneyer *et al. BMC Plant Biology* 2011, **11**:131
http://www.biomedcentral.com/1471-2229/11/131

various stress conditions, different plant tissues and developmental stages. Rye-specific sequences e.g. related to stress tolerance were generated in the present study which are indispensable for functional genomic studies in rye. Second, we reduced the complexity of the transcriptome by cDNA normalization prior to sequencing. cDNA normalization lead to a significant increase in transcriptome sequencing efficiency by equalizing the representation of high, medium and rarely expressed transcripts in the cDNA population [50-52]. Since many transcripts are temporally and/or spatially expressed during plant development, RNA pooled from different tissues at different developmental stages ensured the coverage of temporal- and spatial-specific transcripts.

Linking rye to grass genome sequence resources

To assess, how much of the rye transcriptome is represented by the established EST resource, we compared the Sce_Assembly03 sequences to currently available grass genome, flcDNA, and EST sequences. Generally, the number of sequences with significant BlastX hit in public databases was higher for multi-line contigs than for single-line contigs. This finding is in line with results of Schafleitner et al. [53] who compared EST sequences of sweet potato (*Ipomea batatas*) with sequences contained in the UniRef100 protein database.

The overall gene content across the grass subfamilies *Ehrhartoideae* (rice), *Panicoideae* (maize, sorghum), and *Pooideae* [6] is in a similar range. A total of 25,532 protein coding gene loci were found for *Brachypodium* [6] which is in line with rice [RAP2, 28,236 protein coding gene loci, [32]], maize [ZmB73_v5b.60, 39,656 protein coding loci, [7]], and sorghum [v1.4, 27,640 protein coding gene loci, [5]]. Due to a close evolutionary relationship with these model genomes a pronounced overlap with rye transcripts was expected. The comparison of the Sce_Assembly03 against flcDNA, EST, and genomic sequences revealed a higher homology to barley, *Brachypodium*, and wheat than to maize, rice, and sorghum which was expected, as rye is phylogenetically more closely related to other members of the *Pooideae* than to maize, rice, and sorghum [54,55]. The GO annotation analysis reveals that a broad spectrum of genes was sampled in our normalized cDNA pool from multiple tissues and developmental stages. The large number of reads generated by 454 sequencing entails a substantial gain at the level of gene discovery which provides a valuable resource for forward and reverse genetics approaches in rye as well as for comparative gene analyses. A significant fraction of multi-line contigs (31%) gave no hits with the public grass sequence resources. In part this finding can be attributed to species specific and tribe specific genes and gene families. The *Pooideae* contain 265 subfamily-specific gene families leading to subfamily-specific Blast hits [6]. Given our stringent BlastX/tBlastX cut-off value of > 70% sequence identity, non-conserved and non-coding sequences such as 3'- or 5'- untranslated regions and non-coding RNAs are assumed to contribute to the fraction that lacks homology with other grass species. Around 2% of all rye 454 reads revealed hits to the MIPS Repeat Element database [36], suggesting that transcriptional activity of retrotransposons contributed to the sampled RNA pool. Transcriptome sequencing in two rice subspecies detected alternative splicing patterns in about half of the rice genes and more than 15,000 novel transcriptional active regions of which more than half had no homolog in public protein data [56]. This might suggest that the rye EST resource contains rare, tissue-specific and/or stress-related transcripts that are not represented in sequence resources of the closely related species wheat and barley despite their extensive EST resources. It is anticipated that rye transcriptome sequence analysis will greatly benefit from a reference genome sequence for a member of the *Triticeae* family. Whole genome sequencing is in progress for barley [49,57] and wheat [58] and exploratory BAC end sequencing of rye 1RS-specific BAC libraries [59] has been reported. *In silico* mapping of rye ESTs to the model genome of *Brachypodium* revealed an even distribution of rye transcripts when anchored to their *Brachypodium* homologs. The large extent of synteny between grass genomes will facilitate the construction of a virtual gene map of rye representing the ancestral gene scaffold. Genetic mapping of the SNPs represented on the Rye5K array and of SSRs developed from our rye ESTs is underway and will lead to fine-scale comparative maps between rye and other grasses. A fully annotated genome sequence for rye is still out of reach due to the complexity and highly repetitive nature of the rye genome. However, with the tools established in our study, rye catches up with other grass genome resources and a far more detailed glimpse into the rye genome and its evolution will be possible.

Molecular toolbox for rye

Sequence information of the five rye inbred lines was used to detect sequence variation that was transferred into more than 1,300 SSRs and about 5,000 SNPs. Molecular markers have been developed for a range of crop species and play an essential role in modern plant breeding. They have been used to monitor DNA sequence diversity within and among species, to identify genes responsible for desired traits, to disclose sources of genetic variation that allow for the production of new varieties by introducing favorable traits from landraces and related grass species, and to manage backcrossing programs [60]. Together with amplified fragment length polymorphisms (AFLPs), SSRs are currently the most

popular marker system in cereals. They have been developed for major crop plants including cereals and when applied in breeding programs this marker system is predicted to lead to accelerated progress [61]. Currently, the availability of public rye SSRs is very limited. Our resource significantly increases this marker resource that might facilitate the assessment of genetic variability and the estimation of genetic distances between populations. Besides SSRs the marker system receiving the greatest attention nowadays are SNPs [62]. SNPs have shown huge potential in highly efficient fingerprinting, genetic map construction, marker assisted selection as well as population and evolutionary genetics. The Rye5K SNP array provides a powerful new resource for large-scale genotyping in molecular and genome-centric research in rye. Recently whole-genome genotyping arrays became available for crops and livestock and are used for genome-wide association studies and to investigate genetic variation [e.g. [63]]. In a pilot experiment, we analyzed 59 rye inbred lines including the five lines used for sequencing with the Rye5K SNP array to estimate the degree of residual heterozygosity. Theoretical expectation after two, three or six cycles of selfing is about 12.5%, 6.3%, and 1.6%, respectively. Genotyping of these 59 lines using the Rye5K array showed that the degree of heterozygosity significantly (p-value < 0.05) exceeds this theoretical expectation. This might be in part explained by the allogamous behaviour of rye resulting in remaining heterozygosity [64]. Despite forced selfing during inbred line production some degree of cross-pollination cannot be excluded as the seed was produced as single-ear progenies in a commercial breeding program. The lower levels of residual heterozygosity observed for the seed parent pool is in agreement with the higher advanced selfing generations in rye seed parent lines (P. Wilde, personal communication). A detailed analysis of sequences that remained heterozygous indicated sequences belonging to large gene families, such as transferases and hydroxylases. Detection of SNPs in paralogs or members of gene families may mimic a substantial part of the detected heterozygosity, thus leading to an overestimation of the true remaining heterozygosity in the rye inbred lines.

Conclusions

In conclusion, the Sce_Assembly03 provides a new and comprehensive EST resource that integrates rye in the comparative analysis between small grain cereals. The Rye5K SNP array allows the analysis of large sets of individuals to obtain genotyping data for association studies, estimating linkage disequilibrium, and population genetic approaches. Our genomic resources comprise 115,400 EST sequences, 1,385 SSRs, more than 5,000 SNPs, and the Rye5K SNP array for large-scale genotyping that will improve and promote genetic and genomic research as well as genome-based breeding in rye.

Additional material

Additional file 1: Set of plant tissues for RNA extraction. RNA of each rye inbred line was extracted from plant tissues exposed to various stress treatments and harvested at different developmental stages.

Additional file 2: Establishment and description of the Sce_Assembly02 generated for *in silico* SNP mining. The Sce_Assembly02 was performed in three steps using the MIRA assembler V2.9 on integrated standard settings: Firstly, raw sequence reads surpassed a quality filtering process where 454 sequencing adapter and cDNA synthesis primer sequences as well as low quality reads were removed. Secondly, the cleaned and trimmed sequence reads were subjected to a line-specific assembly where reads of each inbred line were aligned in a separate assembly run. Non-aligned reads in the line-specific assemblies, i.e. singletons, were rejected. Thirdly, those reads that merged into contigs in the line-specific assemblies were moved further to the Sce_Assembly02 starting again with the cleaned and trimmed reads, but now from all five inbred lines. This strategy resulted in contig sequences that were used for SNP detection and subsequently for the design of the high-throughput genotyping SNP array. With regard to SNP discovery this assembly allowed the deduction of critical information about the SNP position like allele coverage.

Additional file 3: GigaBayes parameters. Only parameters different from GigaBayes program default settings are listed.

Additional file 4: Association of multi-line and single-line contigs of the Sce_Assembly03 to the *Brachypodium* chromosomes Bd1 to Bd5. The four heatmaps per chromosome are depicting the density of *Brachypodium* genes, homologous rye sequences, contigs represented on the Rye5k SNP array, and SNPs that were heterozygous among 59 rye inbred lines (from top to bottom) by going along the *Brachypodium* chromosomes in a sliding window with 0.5 Mb window size and a 0.1 Mb shift and determining for each window the number and percent bp coverage of the respective tagged genes. The density values were corrected for the number of Ns per window, if the N content exceeded 60% the value was set to zero and drawn in white color. The number was extrapolated to number per Mb to facilitate comparisons. The heatmaps were created from density values using the Python pylab module in combination with the jet colormap (low to high values from blue to red). Minimum, maximum, and mean number of genes/Mb in *Brachypodium* and hits/Mb in rye, respectively, were given on the left of each map. The ruler on top gives the chromosome length in Mb.

Additional file 5: GO categories found in the Sce_Assembly03 multi-line and single-line contig sequences on Blast2GO level 2. Categories with an occurrence less than 0.05% were summarized in "others".

Additional file 6: SSR motifs detected in 338,536 contigs of the five line-specific assemblies. Mononucleotide repeat motifs were discarded. Mixed motifs describe two close SSR motifs which are separated by less than 100 bp.

Additional file 7: Description of the Rye5K SNP array. SNP containing contigs represented on the Rye5K SNP array were listed including candidate SNP position, probe design sequences provided to Illumina Inc. (San Diego, USA), and GO annotations.

Additional file 8: Observed residual heterozygosity of 54 rye inbred lines representing the two heterotic pools. Heterozygosity was calculated based on genotyping data obtained with the Rye5K SNP array. Lines from the pollen parent pool were in generations F_3 to F_4, lines from the seed parent pool were in generation F_6.

Acknowledgement
We thank Fritz Thümmler (vertis AG, Freising, Germany) for synthesizing and normalizing the cDNA samples, KWS LOCHOW GMBH for providing seed

and DNA samples, and Christof Pietsch for his initial work on the SNP discovery pipeline. This work was supported by a grant [0315063A to E.B., 0315063B to N.S., 0315063C to K.M.] in the framework of the initiative 'GABI-Future' of the German Ministry of Education and Research (BMBF).

Author details
[1]Plant Breeding, Technische Universität München, Centre of Life and Food Sciences Weihenstephan, 85354 Freising, Germany. [2]Bioinformatics and Information Technology, Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Gatersleben, Germany. [3]MIPS/IBIS, Institute for Bioinformatics and Systems Biology, Helmholtz Centre Munich, German Research Centre for Environmental Health (GmbH), 85764 Neuherberg, Germany. [4]Genome Diversity, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany. [5]Genome Analysis, Leibniz Institute for Age Research, Fritz-Lipmann-Institute (FLI), 07745 Jena, Germany.

Authors' contributions
GH prepared the sequencing samples, participated in the bioinformatic analyses, conducted the genotyping, and evaluated the genotyping data. TS, MM, and US carried out the processing and assembly of 454 reads and gave the descriptive statistics for them. KFXM and MS performed the BLAST analyses, functional annotations, and sequence comparisons along the *Brachypodium* chromosomes. NS and RZ developed and examined the SSR markers. EB, GH, and TS developed the Rye5k SNP array. CCS, EB, KFXM, NS, and US designed the study. EB, GH, MS, RZ, and TS drafted the manuscript. All authors read and approved the final manuscript.

Received: 16 February 2011 Accepted: 28 September 2011
Published: 28 September 2011

References

1. Limin AE, Fowler DB: **Cold hardiness of forage grasses grown on the canadian prairies.** *Can J Plant Sci* 1987, **67(4)**:1111-1115.
2. Ko JM, Seo BB, Suh DY, Do GS, Park DS, Kwack YH: **Production of a new wheat line possessing the 1BL.1RS wheat-rye translocation derived from Korean rye cultivar Paldanghomil.** *Theor Appl Genet* 2002, **104(2-3)**:171-176.
3. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, *et al*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*).** *Science* 2002, **296(5565)**:92-100.
4. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, *et al*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*).** *Science* 2002, **296(5565)**:79-92.
5. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al*: **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature* 2009, **457(7229)**:551-556.
6. The International *Brachypodium* Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon*.** *Nature* 2010, **463(7282)**:763-768.
7. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, *et al*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326(5956)**:1112-1115.
8. Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R: **Plant genome size estimation by flow cytometry: Inter-laboratory comparison.** *Ann Bot* 1998, **82(suppl 1)**:17-26.
9. Flavell RB, Bennett MD, Smith JB, Smith DB: **Genome size and the proportion of repeated nucleotide sequence DNA in plants.** *Biochemical Genetics* 1974, **12(4)**:257-269.
10. Shi B, Collins N-C, Langridge P, Gustafson J: **Construction of a rye cv. Blanco BAC library, and progress towards cloning the rye Alt3 aluminium tolerance gene.** *Vortr Pflanzenzuchtg* 2007, **71**:205-209.
11. Hackauf B, Rudd S, van der Voort JR, Miedaner T, Wehling P: **Comparative mapping of DNA sequences in rye (*Secale cereale* L.) in relation to the rice genome.** *Theor Appl Genet* 2009, **118(2)**:371-384.
12. Khlestkina EK, Than MH, Pestsova EG, Roder MS, Malyshev SV, Korzun V, Borner A: **Mapping of 99 new microsatellite-derived loci in rye (*Secale cereale* L.) including 39 expressed sequence tags.** *Theor Appl Genet* 2004, **109(4)**:725-732.
13. Korzun V, Malyshev S, Voylokov AV, Börner A: **A genetic map of rye (*Secale cereale* L.) combining RFLP, isozyme, protein, microsatellite and gene loci.** *Theor Appl Genet* 2001, **102(5)**:709-717.
14. Ma XF, Wanous MK, Houchins K, Milla MAR, Goicoechea PG, Wang Z, Xie M, Gustafson JP: **Molecular linkage mapping in rye (*Secale cereale* L.).** *Theor Appl Genet* 2001, **102(4)**:517-523.
15. Senft P, Wricke G: **An extended genetic map of rye (*Secale cereale* L.).** *Plant Breeding* 1996, **115(6)**:508-510.
16. Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11(1)**:31-46.
17. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10**:219.
18. Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S: ***De novo* sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis.** *BMC Genomics* 2010, **11**:262.
19. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, *et al*: **Whole-genome sequencing and variant discovery in *C. elegans*.** *Nat Methods* 2008, **5(2)**:183-188.
20. Torres TT, Metta M, Ottenwalder B, Schlotterer C: **Gene expression profiling by massively parallel sequencing.** *Genome Res* 2008, **18(1)**:172-177.
21. Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N: **A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley.** *Plant J* 2009.
22. Lister R, Ecker JR: **Finding the fifth base: genome-wide sequencing of cytosine methylation.** *Genome Res* 2009, **19(6)**:959-966.
23. Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M: **High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome.** *BMC Genomics* 2008, **9(1)**:312.
24. Korzun V, Malyshev S, Kartel N, Westermann T, Weber WE, Börner A: **A genetic linkage map of rye (*Secale cereale* L.).** *Theor Appl Genet* 1998, **96(2)**:203-208.
25. Oztur ZN, Talame V, Deyholos M, Michalowski CB, Galbraith DW, Gozukirmizi N, Tuberosa R, Bohnert HJ: **Monitoring large-scale changes in transcript abundance in drought- and salt-stressed barley.** *Plant Mol Biol* 2002, **48(5-6)**:551-573.
26. Kumar S, Blaxter ML: **Comparing *de novo* assemblers for 454 transcriptome data.** *BMC Genomics* 2010, **11**:571.
27. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14(6)**:1147-1159.
28. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, *et al*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437(7057)**:376-380.
29. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9(9)**:868-877.
30. Riano-Pachón DM, Nagel A, Neigenfind J, Wagner R, Basekow R, Weber E, Mueller-Roeber B, Diehl S, Kersten B: **GabiPD: the GABI primary database - a plant integrative 'omics' database.** *Nucleic Acids Res* 2009, , **37 Database**: D954-959.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3)**:403-410.
32. Rice Annotation Project: **The Rice Annotation Project Database (RAP-DB): 2008 update.** *Nucleic Acids Res* 2008, **36(suppl_1)**:D1028-1033.
33. Matsumoto T, Tanaka T, Sakai H, Amano N, Kanamori H, Kurita K, Kikuta A, Kamiya K, Yamamoto M, Ikawa H, *et al*: **Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries.** *Plant Physiology* 2011, **156(1)**:20-28.
34. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K: **TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics.** *Plant Physiology* 2009, **150(3)**:1135-1146.
35. Jurka J: **Repeats in genomic DNA: mining and meaning.** *Curr Opin Struct Biol* 1998, **8(3)**:333-337.
36. Spannagl M, Noubibou O, Haase D, Yang L, Gundlach H, Hindemitt T, Klee K, Haberer G, Schoof H, Mayer KF: **MIPSPlantsDB–plant database**

resource for integrative and comparative plant genome research. *Nucleic Acids Res* 2007, , **35 Database:** D834-840.
37. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Comput Sci Eng* 2007, **9(3)**:90-95.
38. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21(18)**:3674-3676.
39. Thiel T, Michalek W, Varshney RK, Graner A: **Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.).** *Theor Appl Genet* 2003, **106(3)**:411-422.
40. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
41. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet* 1999, **23(4)**:452-456.
42. Huang W, Marth G: **EagleView: a genome assembly viewer for next-generation sequencing technologies.** *Genome Res* 2008, **18(9)**:1538-1543.
43. R Development Core Team: **R: A language and environment for statistical computing.** *R Foundation for Statistical Computing, Vienna, Austria* 2004 [http://www.R-project.org], ISBN 3-900051-07-0.
44. Zagrobelny M, Scheibye-Alsing K, Jensen NB, Moller BL, Gorodkin J, Bak S: **454 pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides.** *BMC Genomics* 2009, **10**:574.
45. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21(9)**:1859-1875.
46. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18(11)**:1851-1858.
47. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, *et al*: **The genome of the cucumber, *Cucumis sativus* L.** *Nat Genet* 2009, **41(12)**:1275-1281.
48. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, *et al*: **The sequence and *de novo* assembly of the giant panda genome.** *Nature* 2010, **463(7279)**:311-317.
49. Mayer KF, Taudien S, Martis M, Simkova H, Suchankova P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, *et al*: **Gene content and virtual gene order of barley chromosome 1H.** *Plant Physiol* 2009, **151(2)**:496-505.
50. Coche T, Dewez M: **Reducing bias in cDNA sequence representation by molecular selection.** *Nucleic Acids Res* 1994, **22(21)**:4545-4546.
51. Emrich SJ, Barbazuk WB, Li L, Schnable PS: **Gene discovery and annotation using LCM-454 transcriptome sequencing.** *Genome Res* 2007, **17(1)**:69-73.
52. Patanjali SR, Parimoo S, Weissman SM: **Construction of a uniform-abundance (normalized) cDNA library.** *Proc Natl Acad Sci USA* 1991, **88(5)**:1943-1947.
53. Schafleitner R, Tincopa LR, Palomino O, Rossel G, Robles RF, Alagon R, Rivera C, Quispe C, Rojas L, Pacheco JA, *et al*: **A sweet potato gene index established by de novo assembly of pyrosequencing and Sanger sequences and mining for gene-based microsatellite markers.** *BMC Genomics* 2010, **11**:604.
54. Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J: **The 'inner circle' of the cereal genomes.** *Curr Opin Plant Biol* 2009, **12(2)**:119-125.
55. Gaut BS: **Evolutionary dynamics of grass genomes.** *New Phytologist* 2002, **154(1)**:15-28.
56. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Huang X, *et al*: **Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq.** *Genome Res* 2010, **20(9)**:1238-1249.
57. Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, *et al*: ***De novo* 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley.** *BMC Genomics* 2009, **10**:547.
58. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, *et al*: **A physical map of the 1-gigabase bread wheat chromosome 3B.** *Science* 2008, **322(5898)**:101-104.
59. Bartos J, Paux E, Kofler R, Havrankova M, Kopecky D, Suchankova P, Safar J, Simkova H, Town C, Lelley T, *et al*: **A first survey of the rye (*Secale cereale*) genome composition through BAC end sequencing of the short arm of chromosome 1R.** *BMC Plant Biol* 2008, **8(1)**:95.
60. Korzun V: **Molecular markers and their application in cereals breeding.** In *Proceedings of the workshop "Marker assisted selection: A fast track to increase genetic gain in plant and animal breeding?": 17-18 October 2003; University of Turin, Italy* Edited by: Lanteri S 2003, 18-22, Electronic forum on biotechnology in food and agriculture.
61. Gupta PK, Varshney RK, Sharma PC, Ramesh B: **Molecular markers and their applications in wheat breeding.** *Plant Breeding* 1999, **118(5)**:369-390.
62. Schulman AH: **Molecular markers to assess genetic diversity.** *Euphytica* 2007, **158(3)**:313-321.
63. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, *et al*: **Development and Characterization of a High Density SNP Genotyping Assay for Cattle.** *PLoS ONE* 2009, **4(4)**:e5350.
64. Adams MW, Shank DB: **The relationship of heterozygosity to homeostasis in maize hybrids.** *Genetics* 1959, **44(5)**:777-786.

doi:10.1186/1471-2229-11-131
Cite this article as: Haseneyer *et al.*: From RNA-seq to large-scale genotyping - genomics resources for rye (*Secale cereale* L.). *BMC Plant Biology* 2011 **11**:131.

PLOS | ONE

RESEARCH ARTICLE

Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition

Moses M. Muraya[1,2☯], Thomas Schmutzer[1☯*], Chris Ulpinnis[1], Uwe Scholz[1], Thomas Altmann[1]

1 Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstraße 3, D-06466, Stadt Seeland, Germany, **2** Department of Plant Science, Chuka University, P.O. Box, 109–60400, Chuka, Kenya

☯ These authors contributed equally to this work.
* schmutzr@ipk-gatersleben.de

Citation: Muraya MM, Schmutzer T, Ulpinnis C, Scholz U, Altmann T (2015) Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition. PLoS ONE 10(7): e0132120. doi:10.1371/journal.pone.0132120

Editor: Lewis Lukens, University of Guelph, CANADA

Received: January 20, 2015

Accepted: June 10, 2015

Published: July 7, 2015

Data Availability Statement: All sequence data has been submitted to EMBL/ENA read archive under the project ID 'PRJEB5496'.

Funding: The KBBE-CornFed project and its research leading to these results have received funding by the German Federal Ministry of Education and Research (FKZ 0315461C); www.bmbf.de. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

A major goal of maize genomic research is to identify sequence polymorphisms responsible for phenotypic variation in traits of economic importance. Large-scale detection of sequence variation is critical for linking genes, or genomic regions, to phenotypes. However, due to its size and complexity, it remains expensive to generate whole genome sequences of sufficient coverage for divergent maize lines, even with access to next generation sequencing (NGS) technology. Because methods involving reduction of genome complexity, such as genotyping-by-sequencing (GBS), assess only a limited fraction of sequence variation, targeted sequencing of selected genomic loci offers an attractive alternative. We therefore designed a sequence capture assay to target 29 Mb genomic regions and surveyed a total of 4,648 genes possibly affecting biomass production in 21 diverse inbred maize lines (7 flints, 14 dents). Captured and enriched genomic DNA was sequenced using the 454 NGS platform to 19.6-fold average depth coverage, and a broad evaluation of read alignment and variant calling methods was performed to select optimal procedures for variant discovery. Sequence alignment with the B73 reference and *de novo* assembly identified 383,145 putative single nucleotide polymorphisms (SNPs), of which 42,685 were non-synonymous alterations and 7,139 caused frameshifts. Presence/absence variation (PAV) of genes was also detected. We found that substantial sequence variation exists among genomic regions targeted in this study, which was particularly evident within coding regions. This diversification has the potential to broaden functional diversity and generate phenotypic variation that may lead to new adaptations and the modification of important agronomic traits. Further, annotated SNPs identified here will serve as useful genetic tools and as candidates in searches for phenotype-altering DNA variation. In summary, we demonstrated that sequencing of captured DNA is a powerful approach for variant discovery in maize genes.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Uncovering genotype-phenotype associations is one of the central goals in a path towards plant improvement, and this requires the accurate detection of different types of genomic variation. The *Zea mays* subsp. *mays*, commonly referred to as maize, reference genome sequence [1] provides a major foundation for maize molecular genetics. However, accurately linking genomic variation to the expression of certain traits requires the systematic investigation and knowledge of the entire spectrum of DNA sequence diversity, including single nucleotide polymorphisms (SNPs), insertions/deletions (INDELs), copy number variation (CNV), and presence/absence variation (PAV), as well as the frequency with which they occur in certain populations. Substantial progress has been made towards this goal in maize, and the insight gleaned from the sequence and structural variation identified in this organism [2–4] has expanded our knowledge of maize evolution and biology and stimulated genome research. Furthermore, the use of known large-scale sequence variation information to develop a comprehensive (50k) SNP genotyping array [5] has demonstrated usefulness in genome-wide association studies (GWAS) [6].

Maize, one of the most important crops for human food and livestock feeds, has a large and complex 2.365 Gbp genome, only 7.5% of which is predicted to encode genes [7]. Maize exhibits very high levels of both phenotypic and genetic variation, with SNP frequencies among maize inbreds higher than those found between humans and chimpanzees [8]. This high level of genetic variation in maize is also manifested in its large pan-genome [9]. Substantial gains in traits of interest have been made through the selection of individuals for breeding based on their phenotypes, or their pedigree. More recently, genomics technologies, such as SNP typing, have been used to select individuals based on their genetic makeup. Targeted SNP genotyping technology has also enabled successful GWAS in maize [5,6,10–14]. However, this has known disadvantages, including ascertainment bias, since the SNPs used in these studies were chosen to exceed a minimum frequency of the rare allele, and SNP markers were selected from a limited number of divergent sources. Thus, the identification of very rare causal mutations might be complicated due to the failure of disequilibrium detection between casual mutations and typed SNPs [15,16]. In contrast, targeted sequencing of individual genomes would be expected to alleviate the polymorphism ascertainment bias. This would support the detection of rare functional variants and allow the discovery of complete haplotypes of genes, as well as CNV and PAV, both of which have recently been identified as an important sources of genomic variation [2–4], when a high coverage depth is attained. With the ever decreasing costs of sequencing and the advances in sequence capture technologies, approaches have been developed to catalogue genomic sequences, CNV, PAV, and rare variants (as opposed to earlier methods that were biased towards common variants) in maize. These can provide key resources for breeding initiatives aimed at mitigating the challenges of increased global demand for food, feed, fiber, and fuel.

As a step towards deciphering the genetic basis underlying important trait variation in maize, genes that are known or assumed to be involved in various aspects of plant growth, biomass production, or composition can be captured and sequenced in a wide panel of diverse maize inbred lines. This will yield a large inventory of genomic variation, which can be used to link genes, or genomic regions, with corresponding phenotypes. To this end, readily available sequence capture and next generation sequencing (NGS) technologies can be used. Unlike the whole genome sequencing that permits deep sequence coverage for only a small number of individuals, sequencing of captured target sequences will result in an enhanced proportion of sequence reads originating from regions of interest. Enrichment procedures allow the redistribution of sequencing efforts from whole genomes of a small number of genotypes to a

restricted genomic fraction of a larger number of genotypes. The use of the latter approach, coupled with the available maize reference genome sequence [1], provides for a substantial contribution to the in-depth characterization of the genetic variation present in this organism.

In this study, we used sequence capture and 454 pyrosequencing to sequence the DNA of 21 maize inbred lines at 4,648 genes of interest and obtained, on average, a 19.6-fold sequencing depth in our raw data. High quality reads were aligned to the maize B73 reference genome sequence, and additionally assembled *de novo* to account for target genomic regions that are absent from the reference sequence. In order to detect the most comprehensive set of reliable SNPs in our maize collection and to determine the optimal variation calling method for our data, a wide range of read alignment and SNP calling tools were evaluated. Optimization of variant calling has previously been attempted by either finding the best alignment tool or by evaluating variant callers [17]. Here, we combined both approaches in order to obtain a broader perspective, and to evaluate how the read alignment procedure impacts the variant calling approach. We thus extended the collection of tools used in recent comparisons [18–20] and further included open source tools, as well as commercial SNP caller. The optimal SNP set was then used to investigate: (1) the level of sequence polymorphism in captured genes across the 21 maize inbreds, (2) the pattern of SNP distribution among the inbred lines and their functional annotation relative to B73, and (3) the pattern of gene variation and presence/absence genes in the studied inbred lines.

Results

Array design, sequence capture optimization, and assessment of capture efficiency

A high-density (2.1 million) oligonucleotide sequence capture microarray was designed using the B73 genome sequences. Captured target regions were selected ranging in sizes from 58–4,240 bp, with an average probe length of 75 bp. Probe selection settings allowed for up to 5-matches when aligned to the B73 RefGen_v1, and these probes were classified based on repetitiveness and locations relative to predicted genes (see Materials and Methods).

To modify and optimize available sequence capture protocols, a set of four customized qPCR loci was employed to estimate relative enrichment and to determine whether a capture was successful prior to sequencing (see Materials and Methods). We first tested capture efficiency of B73 DNA, and obtained high enrichments for all four control loci, ranging from 643- to 990-fold, with a mean of 749-fold (S1 Table). The enrichment of captured DNA from the remaining 20 inbred lines was then analyzed using the same control loci, and a high enrichment was found for all captured DNAs, ranging from 623- to 755-fold mean enrichment. NimbleGen recommends at least a 300-fold enrichment before committing sample libraries to the expensive and/or time-consuming downstream applications. In order to enhance the robustness and reliability of our DNA library quantification, we developed a method based on qPCR. Because our capture libraries, on average, contain 700 bp fragments, we used a plasmid fragment that results in a 725 bp PCR product when ligated to two 454 adaptors. The resulting PCR product, being of known quantity, was used as a standard for quantification of the captured DNA libraries. This leads to a more precise quantification of enriched captured DNA and minimizes the variation in cluster density or template-to-bead ratio, thus reducing the failure of GS FLX Titanium emulsion PCR (emPCR) preparation and subsequent sequencing. For example, using the standard DNA quantification protocol, the captured 454 sequence output for B106 and Mo17 was 50 and 37 Mbp, respectively. Whereas, after utilizing the optimized captured DNA quantification protocol, a sequence output of 235 Mbp and 312 Mbp, respectively, for B106 and Mo17 was achieved.

454 pyrosequencing

A total of 17,766,241 sequence reads, with an average read length of 363 bp, was generated, yielding more than 6.4 Gbp of sequence data (Table 1). The resulting average sequencing depth was estimated at 19.6-fold for the target regions. Among these reads, 10.6 million (60.0%) passed the quality trimming process and attained an average read length of 286 bp (78.8%). Upon quality trimming, the sequence depth declined to 9.3-fold (on average), which is still suitable for SNP detection (Table 1).

Evaluation of alignment methods and mapping results

The combined set of reads was aligned to the B73 reference genome (AGV3) using seven different read alignment tools (Table 2, see Materials and Methods), and the total fraction of mapped quality trimmed reads ranged from 95.66% (Bowtie2) to 99.81% (Stampy). On average 98.7% of the reads were aligned to the B73 reference. On average, 41.32% of the quality trimmed reads were mapped to the target sequence, ranging from 35.57% (Bowtie2) to 41.81% (BWA--MEM). However, the mapped sequence depth variation was minimal for all aligners (Table 2). In addition, an extended series of 441 independent read alignments was performed for all 21 inbred lines, using three different parameter settings and was evaluated to reveal the most reliable setup (Fig 1A, see Materials and Methods). The agreement between the different read alignment methods was analyzed by calculating those reads that were mapped by the majority of tools, as well as the number of reads mapped by at least one other alignment method (S1 Fig). For all seven evaluated tools the constructed read alignments reached high quality.

However, our results revealed a sub optimal performance of the tools NGM and Smalt using standard parameter settings. Furthermore, our results indicate that Bowtie2, being one of the most widely used read alignment tools for Illumina sequences, performed less optimal utilizing 454 sequence reads. Thus, we emphasize that read alignment tools and parameter settings should be carefully assessed. Finally, we conclude that BWA-MEM performed best across all genotypes resulting in high confidence of sequence alignments (99.99% of aligned reads were in agreement with another alignment method) and with premier exactness in terms of total aligned reads (99.74%) and reads aligned on target (41.81%).

To evaluate the impact of read alignment on the quality of variant calling, the results of all seven read alignment tools (BAM files) were processed by the all variant calling methods on a randomly selected inbred line (NC358). For this evaluation, we analyzed 504 possible combinations of read alignment and variant calling tools (each individual method was applied in the three parameter settings). Ten incompatible combinations were spotted, which lead to failures (no results obtained) and the overview of all true positive sites detected in each of the individual approaches is depicted in Fig 1B. To determine the impact of read mapping, we compared the merits of each variant caller using different read alignment results as inputs. Using the 504 combinations of read alignment and variant calling methods, we analyzed how the number of successfully detected true positive sites varies with application of different read alignment methods. The observed variability (detected range of true positive sites for a particular variant caller) was, on average, 16.41% (maximum 26.15%) per variant calling method (S2 Fig). The analysis was most consistent for CLC, with only about 5% variability, whereas SNVer was the most variable across different read alignment tools. We then checked the quality of the allele calls obtained using the various tool combinations by comparing them to the genotyping data of the maize 50k SNP array (S2 Table), again using inbred line NC358 as an example. An average genotype concordance of 97.76% was observed, ranging between 95.77% and 98.92% for the various tool combinations. In addition to being the best performer with respect to the number of aligned reads, BWA-MEM also performed best in this analysis. Consequently,

Table 1. Statistics of raw and preprocessed sequence data. A total volume of >6.4 Gbp of raw data was sequenced. Each of the 21 maize genotypes is designated by 'D' (Dent) or 'F' (Flint), according to which gene pool the line belongs.

Genotype	Total bases (Mbp)	Mean length (bp)	Raw reads	Quality trimmed reads	%	Base pairs used after trimming	%	Estimated coverage	Estimated coverage (trimmed)
B73 [D]	399	402	992,806	595,595	59.99	186,036,442	46.62	25.6	11.9
F353 [D]	200	397	503,938	324,246	64.34	101,734,580	50.82	12.8	6.5
P068 [D]	251	387	647,422	427,578	66.04	136,393,311	54.37	16.1	8.7
UH007 [F]	445	392	1,134,431	708,594	62.46	220,687,755	49.61	28.5	14.1
B101 [D]	230	294	782,915	433,769	55.40	95,362,651	41.47	14.7	6.1
B102 [D]	283	344	823,149	474,658	57.66	127,937,630	45.20	18.1	8.2
P107 [D]	273	340	804,167	433,932	53.96	112,451,816	41.12	17.5	7.2
B106 [D]	285	319	1,017,225	574,263	56.45	122,507,030	43.00	18.3	7.9
B111 [D]	342	382	895,724	631,430	70.49	195,512,634	57.16	21.9	12.5
F7059 [D]	303	371	817,130	504,308	61.72	147,282,945	48.64	19.4	9.4
Mo17 [D]	349	324	1,076,211	651,146	60.50	164,714,385	47.22	22.4	10.6
Mo24W [D]	308	382	804,525	450,652	56.01	133,649,118	43.43	19.7	8.6
NC358 [D]	340	387	879,208	537,185	61.10	164,251,568	48.31	21.8	10.5
P128 [D]	331	386	855,009	466,105	54.51	139,512,583	42.17	21.2	8.9
DK105 [F]	326	361	902,739	532,245	58.96	149,157,334	45.80	20.9	9.6
EA1070 [F]	337	370	909,460	532,589	58.56	155,772,546	46.26	21.6	10.0
EP1 [F]	343	380	903,222	486,639	53.88	146,530,019	42.68	22.0	9.4
F2 [F]	312	382	816,728	437,189	53.53	130,277,657	41.76	20.0	8.4
F7 [F]	316	387	816,142	515,399	63.15	166,516,436	52.76	20.2	10.7
Lo11 [F]	310	388	799,324	511,863	64.04	164,539,579	53.01	19.9	10.5
PH207 [D]	149	254	584,766	393,234	67.25	80,728,746	54.35	9.5	5.2
sum	6.432		17,766,241	10,622,619		3,041,556,765		412.2	195.0
avg.	306	363	846,011	505,839	60.00	144,836,036	47.41	19.63	9.28

doi:10.1371/journal.pone.0132120.t001

Table 2. Evaluation of seven read alignment methods. In this broad evaluation 21 inbred maize lines were included. Seven independent read alignment methods were utilized in three different parameter settings. For each alignment methods the best parameter setting is shown with respect to the highest number of reads mapped on target.

Read alignment method	# mapped reads *	Mapped reads [%] *	# reads mapped on target (captured regions)	Reads mapped on target [%]	25 pctil of reads mapped on target [%]	75 pctil of reads mapped on target [%]	Reads mapped by majority (>4) of tools [%]	Reference
Bowtie2	9,957,961	95.66	4,182,513	40.18	35.57	45.89	95.48	[21]
BWA MEM	10,359,385	99.52	4,352,690	41.81	36.47	47.84	98.83	[22]
BWA SW	10,275,404	98.71	4,314,371	41.45	35.77	47.79	98.16	[23]
CLC mapper	10,309,608	99.04	4,299,431	41.30	36.35	47.16	98.55	[24]
NextGenMap	10,285,973	98.81	4,306,285	41.37	36.49	47.33	98.38	[25]
Smalt	10,374,995	99.67	4,322,142	41.52	36.50	47.43	98.83	[26]
Stampy	10,389,771	99.81	4,328,226	41.58	36.61	47.45	98.87	[27]

*Number of total reads is 10,409,726

doi:10.1371/journal.pone.0132120.t002

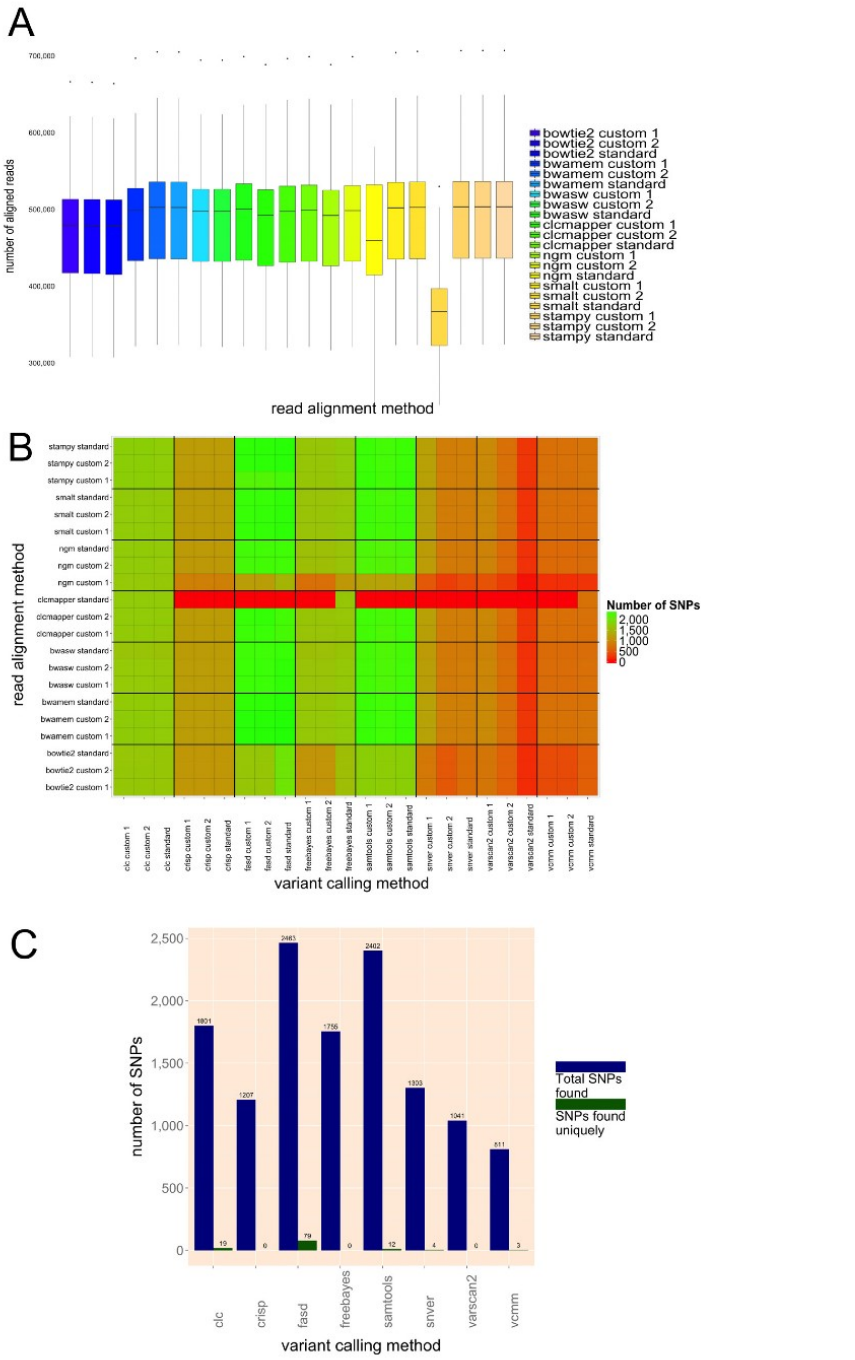

Fig 1. Evaluation of read alignment and variant calling methods. (A) Comprehensive illustration of 441 evaluated read alignment results. Each method is referenced in standard, and in two additional, parameter

settings. The plots show the number of aligned reads, where the range for each bar illustrates the observed variability when different lines were used. **(B)** Heat map depicts the true positive sites in the 50k array. A total of 504 combinations of read alignment and variant calling methods were evaluated to identify recommended or less optimal applications (genotype NC358). **(C)** Variant caller performance compared to the 50k array. The total number of identified SNPs, as well as the number of unique SNPs, is depicted for each of the eight evaluated methods (genotype NC358).

doi:10.1371/journal.pone.0132120.g001

BWA-MEM was selected as the preferred tool for subsequent mapping of our 454 maize sequence data.

De novo assembly of captured sequence reads

Quality trimmed 454 reads, totaling ~3 Gbp of sequence, were used to perform *de novo* assemblies, individually for each inbred line applying Newbler (version 2.6, 454 Life Sciences, Basel, Switzerland) with the default settings. Overall, the number of contigs per inbred line ranged from 8,512 to 19,515, and the contig sizes across all inbred lines ranged from 300 bp to 16,184 bp (S3 Table). A total of 15,461 assembled contigs were obtained for the B73 sequences, with a median size of 903 bp and the largest contig encompassing 8,827 bp; this resulted in 14 Mbp of non-redundant sequence. *De novo* assembly of reads from the other genotypes revealed that some of the targeted gene sequences were not completely covered, resulting in more fragmented assemblies.

Evaluation of variant calling methods

In previous studies, a low concordance of variant calling algorithms was observed when comparing independent variant calling methods [19,20,28]. This can be due to a number of reasons, such as the use of different internal cut-off values, filtering, or variable incorporation of parameters. Further development of variant calling methods is mainly driven by the motivation to implement novel approaches that is faster and more sensitive than concurrent tools. In general, variant calling refers to a Bayesian-based algorithm in order to predict the consensus genotype. Although these are widely used standard methods, different tools incorporate different information to determine the corresponding genotype. Several important metrics exist that describe the quality of a detected SNP; these include 'strand bias', the phred-scaled 'quality score', the neighboring quality score (NQS), and the coverage filter. The usage of post-filtering is often recommended over the use of internal hard cut-off values by many variant calling tools [29,30]. We therefore decided to employ variant calling at low stringency utilizing the default settings.

To estimate the false positive rates of various SNP discovery protocols, a defined set of verified polymorphic positions was compiled by extracting SNPs from several control data sets (50k, GBS, RNAseq, and HapMap2), which are located within our target regions. These were used to assess the sensitivity (S_e) and specificity (S_p) of eight variant calling methods (SAMtools [31], VarScan2 [32], CRISP [33], CLC find_variations [24], FaSD [34], SNVer [35], VCMM [36], and Freebayes [30]). This evaluation involved two approaches, the first of which involved a single inbred line (NC358) for which an in-depth evaluation was performed using various custom, as well as standard, parameter settings for read alignment and variant calling methods (see Material and Methods). The second approach included all 21 inbred lines, applying the standard settings. For the in-depth evaluation, we used the raw variant calling (no filtering) of candidate sites to assess the performance of the detection process without any bias of posterior filtering. Using this method, we observed that the largest number of called variant positions

(VP) was observed for FaSD and SAMtools, whereas CRISP, VarScan2, and VCMM detected lower numbers (Table 3). Further assessment revealed that FaSD and SAMtools display the highest sensitivity in calling true positive sites among all tested control data sets (Fig 1C). However, their settings that allow the calling of a large number of putative VPs come at the cost of precision. This is demonstrated in the F_1-score, which is defined as harmonic mean between sensitivity and precision. For FaSD and SAMtools, we observed average F_1-scores across all control data sets of 0.25 and 0.24, respectively. In contrast CRISP (0.59) and VarScan2 (0.58) had the highest F_1-scores, but at the cost of higher numbers of false negative values.

With the second, broader evaluation approach, the overall tendencies in terms of S_e values were confirmed for the majority of evaluated variant calling methods (Table 3, second part). The highest average sensitivities among all the external control datasets was observed in SAMtools (0.94), FaSD (0.71), and CLC (0.70). This larger dataset (all genotypes) lead to a general increase of the total number of detected variant positions. However, a stringent posterior filtering was applied in order to control for the high number of detected VPs that has already been observed (see Materials and Methods). Consequently, the number of false positive sites decreased, resulting in an increase of the F_1-score, and a subsequent improvement of the overall reliability of the prediction.

This extensive investigation of multiple variant calling methods revealed that that none of the evaluated methods completely captured the complete set of VPs. The low concordance between variant calling methods further highlights the benefit of applying multiple independent tools to obtain a conclusive diversity set. As others have noted [37], VPs detected by multiple tools have a higher validation rate than caller-specific VPs. In that respect, our analysis provides supporting evidence and thus, the inclusion of multiple variant calling methods to improve the validity in the final called SNP set is recommended. We balanced our cutoff to call a specific variant site confident when three independent variant calling methods support the prediction. The cutoff value used in our study was achieved by an analysis of F_1-scores from different settings that revealed a peak of the F_1-score at value of three. With this approach we gained confidence and establish a positive validation, without discarding too many SNP candidates by an over-rigorous setting (S3 Fig). Thus, we demonstrate the advantageous effect of applying multiple variant calling methods.

Large-scale sequence polymorphism discovery

A collection of approximately 4.8 million VPs was called across all 20 non-B73 inbred lines when these were compared against the B73 reference sequence in the initial phase of the discovery process. The inclusion of a pre-filtering step, requiring minimal coverage of at least five reads per VP, led to a global total of 696,665 variants, including ~10% insertions and deletions (INDELs). However, a final list of approximately 42,000 INDELs was obtained after discarding positions overlapping with homopolymers (>7 bp) in the maize reference sequence, which are error prone in 454 sequences [38]. We found that 71% of detected VPs were located on target, while 29% were off-target. Off-target positions were uniformly distributed among all ten maize chromosomes, and several variants were detected in mitochondrial (249) or plastid (536) sequences. Of all SNPs and INDELs called as off-target, 74.9% and 68.2%, respectively, were located in genic regions (off-target captured genes), likely mapping to the paralogous sequences that are closely related to the target genes (Fig 2).

To further increase the reliability of *in silico* VPs prediction, the results of multiple prediction tools were combined for validation. This combinatorial approach required that a VP be predicted by at least three independently performed variant calling methods (S3 Fig). Since different variant calling tools use different quality score scales to measure the significance of a

Table 3. Variant detection performance. Comprehensive overview of eight evaluated variant detection tools. Predicted VPs of each variant caller are compared to the four control data sets (50k, GBS, RNAseq, and HapMap2) in terms of sensitivity ('S_e'), specificity ('S_p'), and the F_1-score ('F_1'). In addition the final set of variants detected in this study (CTD) is showing the proportion each variant caller is capturing.

Variant calling method	#total variant calls	CTD*			50k			GBS			RNAseq			HapMap2		
		S_e	S_p	F_1	S_e	S_p	F_1	S_e	S_p	F_1	S_e	S_p	F_1	S_e	S_p	F_1
	(NC358)															
	total		128,064			2,524			9,608			27,620			300,664	
clc	639,075	0.766	0.972	0.256	0.714	0.972	0.267	0.629	0.972	0.270	0.721	0.972	0.277	0.426	0.975	0.342
CRISP	79,194	0.385	0.998	0.475	0.478	0.998	0.760	0.294	0.998	0.732	0.416	0.998	0.694	0.138	0.999	0.293
FaSD	986,133	0.823	0.955	0.189	0.976	0.955	0.194	0.857	0.955	0.199	0.868	0.955	0.207	0.880	0.964	0.457
freebayes	309,230	0.717	0.988	0.420	0.695	0.988	0.458	0.504	0.989	0.458	0.616	0.989	0.461	0.326	0.991	0.425
SAMtools	1,090,172	0.890	0.950	0.187	0.952	0.950	0.190	0.854	0.950	0.195	0.865	0.950	0.201	0.847	0.959	0.420
SNVer	262,246	0.613	0.990	0.403	0.516	0.990	0.460	0.402	0.990	0.458	0.507	0.990	0.455	0.228	0.991	0.346
VarScan2	115,455	0.499	0.997	0.524	0.412	0.997	0.707	0.280	0.997	0.688	0.392	0.997	0.659	0.151	0.998	0.327
VCMM	121,012	0.413	0.996	0.424	0.321	0.996	0.602	0.204	0.996	0.583	0.310	0.996	0.552	0.109	0.997	0.254
	(read depth >5)															
	total		383,145			6,127			26,488			73,891			517,877	
clc	716,777	0.777	0.979	0.541	0.772	0.979	0.587	0.653	0.979	0.588	0.732	0.980	0.594	0.559	0.984	0.599
CRISP	308,032	0.584	0.996	0.648	0.725	0.996	0.841	0.549	0.996	0.832	0.639	0.996	0.823	0.380	0.999	0.618
FaSD	467,723	0.755	0.991	0.679	0.832	0.991	0.764	0.680	0.991	0.765	0.713	0.992	0.763	0.592	0.997	0.759
freebayes	579,101	0.524	0.981	0.417	0.725	0.981	0.516	0.529	0.981	0.518	0.528	0.982	0.517	0.546	0.989	0.617
SAMtools	904,019	0.939	0.973	0.560	0.985	0.973	0.573	0.946	0.974	0.584	0.927	0.974	0.599	0.910	0.987	0.809
SNVer	201,330	0.524	1.000	0.686	0.567	1.000	0.991	0.396	1.000	0.959	0.477	1.000	0.910	0.288	1.000	0.521
VarScan2	355,679	0.619	0.994	0.642	0.682	0.994	0.799	0.500	0.994	0.789	0.540	0.994	0.771	0.418	0.998	0.640
VCMM	367,675	0.879	0.998	0.897	0.615	0.998	0.952	0.451	0.998	0.936	0.532	0.998	0.911	0.340	0.998	0.645

*'CornFed Target Diversity' (CTD) is the final set of VPs in chromosomes.

doi:10.1371/journal.pone.0132120.t003

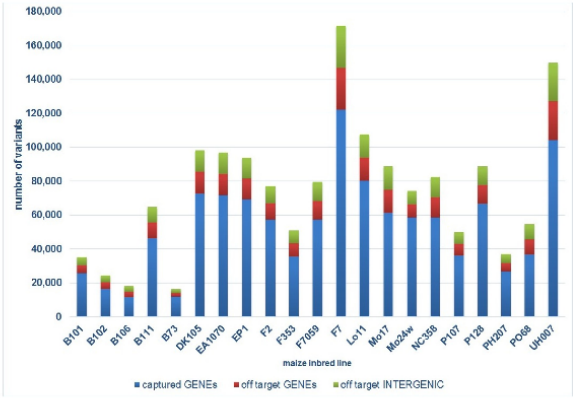

Fig 2. Global diversity classification per genotype. Variant positions in captured genes (on target) identified after basic filtering (at least 5-fold coverage of read depth at SNP position) and off-target regions for different maize inbred lines.

doi:10.1371/journal.pone.0132120.g002

prediction, we normalized the score and applied a generalized minimal threshold of 0.4, where 1.0 is the maximal score of the particular tool. A total of 383,145 variants (including ~91% SNPs and ~9% INDELs) were called across the 20 non-B73 inbreds, with an average of 45,594 SNPs per line (Table 4). In the following, we refer to this variant set as 'CornFed Target Diversity' (CTD), emphasizing that the observed diversity is confined to the investigated set of target genes and closely related sequences across the studied lines. Out of these, only 127 SNPs were called in all non-B73 inbred lines (Table 4). A total of 18,489 insertions and 23,039 deletions were called across theses inbred lines, with global average of 1,925 and 2,469, respectively, per inbred line (Table 4).

Finally, to estimate the variant caller performance, we determined the fraction of the final list of variants (383,145 sites) detected by each variant caller in our data set. This was achieved by looking at the sensitivity (S_e) of each variant calling method in the CTD set (Table 3). According to these criteria, SAMtools (0.94) was ranked first, followed by VCMM (0.88), CLC (0.78), FaSD (0.76), VarScan2 (0.62), CRISP (0.58), Freebayes (0.52), and SNVer (0.52), in the detection of VPs that were ultimately confirmed in our CTD set (Table 3). The most balanced performance was observed for VCMM, which simultaneously displayed a relatively small number of total VPs and an exceptionally high number of true positive sites. Consequently, VCMM (0.90) was ranked first with respect to the F_1-score, followed by SNVer (0.69), FaSD (0.68), CRISP (0.65), VarScan2 (0.64), SAMtools (0.56), CLC (0.54), and Freebayes (0.42).

Using our combinatorial variant calling approach, we achieved a new level of confidence for SNP calling in inbred maize lines. The comparison of this method to the stand-alone application of each single variant caller showed that purely filtering by read depth would result in almost double the number of VPs (681,993). The majority of additional VP candidates (299,971) was observed by only a single tool and in only a single genotype (56.0%). In addition, 24.1% were solely predicted by a single tool, and another 11.2% were observed only in a single genotype. Although these sites might be true rare allelic variants, the probability of erroneous calls is much higher due to very low statistical evidence [42]. The analysis of polymorphic information content (PIC), as defined previously [43], revealed a very low PIC value for removed positions. With a median value of 0.07, only 6,982 sites (2.33%) were characterized with sufficient values (0.2–0.5). Consequently, these indicate a lower reliability, supporting the

Table 4. SNP functional class membership. Re-sequencing and variant calling within 21 maize inbred lines discovered 383,145 high quality candidate SNPs (complete S12 Table). Table displays the five least diverse (green) and five most diverse (blue) inbred lines, also differentiating between the dent ['D'] and flint ['F'] inbred lines. The intersection ($\cap$) was analyzed in four settings (5, 10, 15, and complete), indicating that the variant site is present in at least the respective number of genotypes.

Maize inbred line	B106 [D]	B102 [D]	B101 [D]	PH207 [D]	P107 [D]	DK105 [F]	EA1070 [F]	Lo11 [F]	F7 [F]	UH007 [F]	$\cap$ 5+	$\cap$ 10+	$\cap$ 15 +	$\cap_{complete}$	$\cup_{complete}$
Total variants	11,386	16,190	22,215	22,624	31,602	61,253	63,533	67,983	74,820	96,158	54,212	11,721	2,209	127	383,145
Homozygous	6,308	9,467	13,104	13,907	20,299	40,688	42,465	46,490	49,292	56,986	10,124	1,620	231	1	265,728
Insertions	385	487	849	953	1,234	2,830	2,773	3,392	3,317	3,728	1,958	238	27	0	18,489
Deletions	437	621	1,128	1,062	1,599	3,465	3,496	4,000	4,025	5,547	2,603	402	68	8	23,039
Variants overlapping target region	6,723	10,412	15,847	16,092	22,143	44,418	44,463	47,689	49,992	61,059	10,879	2,201	451	8	259,547
Variants overlapping 50k [5]	174	272	355	367	550	976	907	1,092	1,049	1,121	305	77	16	0	3,822
Variants overlapping GBS [39]	442	582	864	1,043	1,196	2,520	2,373	3,063	2,948	3,120	640	151	27	0	12,432
Variants overlapping RNAseq [40]	1,544	2,624	3,580	3,639	5,155	9,384	9,186	10,508	10,390	11,544	2,566	690	140	1	41,226
Variants overlapping HapMap2 [41]	4,726	7,472	10,452	10,651	16,260	32,009	33,838	35,868	38,136	49,119	8,647	1,594	239	1	190,347
Functional classes															
SNPs in exons	4,755	6,672	9,776	10,204	12,268	24,432	23,091	26,167	27,811	31,372	15,586	3,776	664	13	86,599
Silent mutation	7,635	11,001	14,982	14,984	22,711	44,794	48,033	50,558	55,745	75,227	38,050	7,615	1,428	105	299,126
SNPs in introns	1,577	2,426	3,360	3,329	5,639	11,259	11,673	11,878	12,866	18,427	2,844	461	78	7	70,218
SNPs in UTR	1,844	2,714	3,987	4,638	6,175	13,268	13,967	15,214	15,622	19,050	3,136	523	93	2	89,583
Non-Synonymous SNPs															
non conservative missense	320	448	591	693	690	1,383	1,295	1,416	1,603	1,836	1,229	286	54	0	8,228
conservative missense	1,433	1,863	2,576	2,836	3,103	5,881	5,437	5,968	6,807	7,653	5,388	1,314	217	5	34,457
Synonymous SNP	1,845	2,654	3,699	3,708	4,716	8,263	7,936	9,032	9,561	10,202	8,687	2,115	377	7	41,334
Frame shift	58	95	142	186	153	404	351	438	489	564	282	61	16	1	2,580
SNPs located in splice site															
variant in splice site[a]	116	180	276	262	404	701	675	774	820	998	151	35	6	1	4,131
variant in essential splice site[b]	27	28	56	48	65	132	143	128	165	211	26	6	1	0	931
variant in splice donor site	14	15	22	26	29	44	46	56	63	93	45	10	2	0	364
variant in splice acceptor site	8	11	32	26	31	72	59	71	80	92	55	12	4	0	388
SNP leading to premature terminal codon	38	48	77	65	64	159	141	161	168	216	110	13	5	0	1,170
SNP eliminating terminal codon	3	8	14	7	21	33	26	35	44	37	29	10	6	0	155
Transition[c]	6,957	9,731	12,713	13,214	17,792	33,410	34,826	35,757	40,809	54,148	8,120	1,631	367	13	229,728

(*Continued*)

Table 4. (*Continued*)

Maize inbred line	B106 [D]	B102 [D]	B101 [D]	PH207 [D]	P107 [D]	DK105 [F]	EA1070 [F]	Lo11 [F]	F7 [F]	UH007 [F]	∩ 5+	∩ 10+	∩ 15 +	$\cap_{complete}$	$\cup_{complete}$
Transversion[d]	2,985	4,474	6,040	5,920	8,800	17,138	17,877	19,445	20,992	26,465	4,400	830	171	7	106,399

[a] SNP is located 1–3 bases into an exon or 3–8 bases into an intron

[b] SNP is located in the first two or the last two bases of an intron

[c] transition (A <-> G, C <->T)

[d] transversion (C <-> G, A <-> C, G <-> T, A <-> T)

∩ intersection using the complete SNP data set of 21 genotypes

∪ union using the complete SNP data set of 21 genotypes

doi:10.1371/journal.pone.0132120.t004

decision to discard these positions. An alternative option to the applied combinatorial variant calling is to perform a more stringent filtering at higher read depth. Increasing the applied threshold of read depth resulted in a very substantial reduction of detected sites with a high likelihood of losing meaningful VPs. When a read depth threshold of 10 per site was applied, only 31.6% of the final CTD calls were detected. We conclude that the application of a combinatorial variant calling approach offers a higher reliability than stand-alone methods and thus is a useful option for diversity calling, especially in low-coverage sequencing.

Genetic relationships among the 21 inbred maize lines

The genetic relationships among the 21 inbred maize lines were assessed based on the genetic variation across the re-sequenced genes and the corresponding SNP profiles. SNP profiles clearly differentiated the inbred lines into the two major European gene pools (Dent and Flint, Fig 3). The flints were further differentiated into two groups, with lines originating from Spain in one group (except EP1), and lines originating from France or Germany in the second, larger group. This larger group was further differentiated into three subgroups, representing lines originating from Germany, Spain, and France. The dent lines were further differentiated into three major groups, the first consisting of USA lines, the second containing lines from Germany and France (except PH207, which is a USA material), and a third group consisting of lines from Canada, USA, and France.

STRUCTURE analysis predicted K = 8 as the optimum number of sub-populations, revealing that at least eight distinct groups exist in the studied inbred lines (Fig 4 and S4 Fig). Some groups displayed heterogeneity, comprising a sizable portion of another group; however, most inbred lines originating from the USA clustered together in one group (yellow-filled bar). In

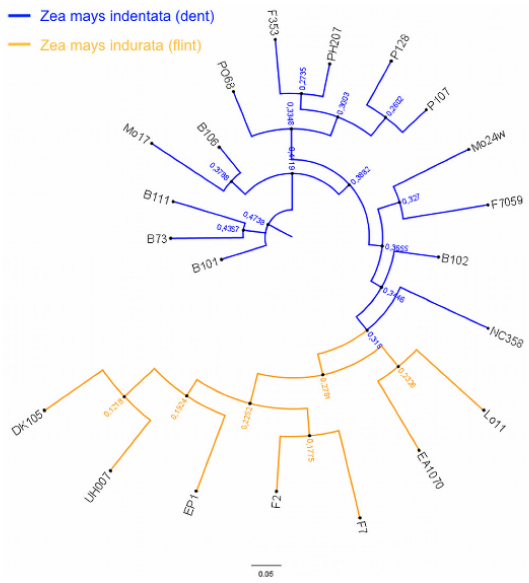

Fig 3. SNP based phylogenetic tree. Phylogenetic tree depicting the SNP distance between 21 maize inbred lines, emphasizing the diversity in this collection. Over 265,000 homozygous SNP calls have been processed to construct this dendrogram.

doi:10.1371/journal.pone.0132120.g003

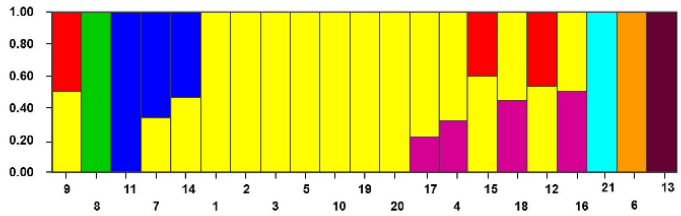

Fig 4. Bar plots of the STRUCTURE analysis. Each of the 21 maize inbred lines is represented by a vertical bar, partitioned into K = 8 colored segments that designate the fraction of each population estimated to belong to the inferred subgroups (Population ID corresponds to S1 Table).

doi:10.1371/journal.pone.0132120.g004

regards to the two main gene pools (Dent and Flint), the majority of flints, except those originating from France, were classified into four distinct groups, each consisting of an individual inbred line. The flints originating from France contained a significant portion of USA dent materials. On the other hand, dent materials (except F7059, originating from France) display heterogeneity, comprising a portion from at least one of the four dent groups.

Functional analysis of SNPs

To exclude most erroneously called VPs from the functional analysis, only the final set of variants comprising 383,145 polymorphic positions (including INDELs), which was validated *in silico* by the combinatorial variant calling approach, was considered. A total of 229,728 and 106,399 SNPs were predicted to be transitions or transversions, respectively (Table 4). Of these, 86,599 (22.6%) were located in exons, with a global average of 11,406 SNPs per inbred line (Table 4). The majority of SNPs (78.1%) represented silent mutations. However, a substantial number (42,685 SNPs) were non-synonymous (nsSNPS), with 8,228 SNPs classified as non-conservative missense and 34,457 SNPs classified as conservative missense. A total of 2,580 INDELs were annotated as frameshift mutations. In addition, 4,131 VPs were detected in splice sites, including 931, 364, and 388 VPs that were classified as variants in the essential splice site, splice donor, and splice acceptor, respectively. We further found 155 SNPs that were predicted to eliminate a terminal stop codon and 1,170 SNPs that inserted a premature stop codon.

To determine the number of novel variants detected in this study, we calculated the overlap between variants detected in this analysis and those listed in previously published maize variant resources. Using this comparative analysis, 3,822 (7.3%) VPs from the 50k array, 12,432 (1.7%) from the GBS, 41,226 (4.4%) from RNAseq, and 190,347 (0.3%) positions from the HapMap2 data set were ascertained to be consistent with those identified here. Thus, the majority (~52%) of *in silico* predicted variants were found to corroborate those available in public data sets. However, a large proportion of variants were newly detected in this study (Fig 5), leading to a set of 185,211 variant positions that have not yet been documented.

Gene diversity and functional annotation of target genes accumulating radical mutations

To measure gene diversity across the 21 inbred lines using the final set of variants (CTD), a heat map illustrating the diversity of captured genes in these lines in comparison to the B73 reference line was constructed (Fig 6). We observed that the majority of genes displayed low diversity in terms of nucleotide changes. However, interesting patterns of divergence between the elite flint and dent inbred lines were observed near the ends of chromosomes in regions of high diversity. Specifically, flint inbred lines displayed high nucleotide variation on

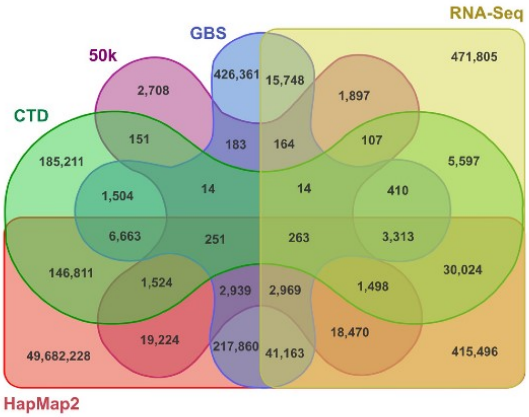

Fig 5. Venn diagram of marker/variants in maize. Illustration of the overlapping intersections between four diversity data sets (50K, GBS, RNAseq, and HapMap2) and the 'CornFed Target Diversity' (CTD).

doi:10.1371/journal.pone.0132120.g005

chromosome 3 at physical position 109 Mbp, while dents had only minor changes here. Conversely, an opposite pattern was observed on chromosome 8 at physical position 114 Mbp, where dent inbreds were found to be more diverse, as compared to B73, than flint inbred lines (S6 Fig).

As stated above, a total of 86,599 unique SNPs was detected in exons (coding SNPs), with an average of 11,762 SNPs per inbred line (S4 Table). The majority (82.0%) of these coding SNPs were classified as rare variants, as they occur in genes affecting five or fewer inbred lines. Only 2,843 coding SNPs were detected in more than half of the inbred lines, 13 of which were found

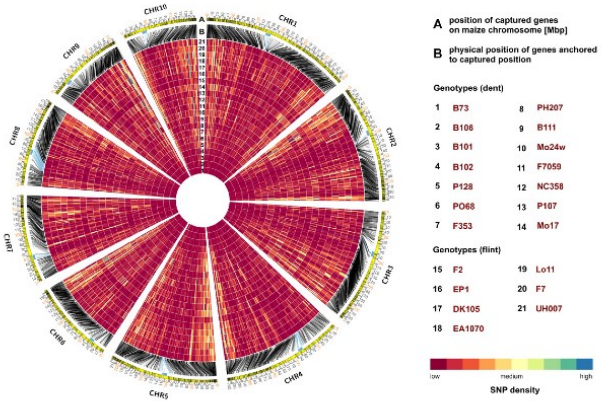

Fig 6. Captured target gene diversity in 21 maize inbred lines for biomass production and composition. (A) The outer circle illustrates, with dark connecting lines, the location of the 4,372 target genes in the 10 maize chromosomes. The inner circles (heat maps 1–21) represent the 21 (ordered in dent and flint inbred lines and subsequently with increasing diversity) inbred lines and their corresponding SNP density, scaled according to the physical position of the target genes in the B73 genome reference **(B)**. In total, 1 million of the detected high quality SNPs are integrated. Putative regions of high diversity are represented in blue, and regions with low accumulations of mutations are colored red. Centromere positions are indicated with light blue bars adjacent to the outer circle.

doi:10.1371/journal.pone.0132120.g006

in all 20 lines. The number of genes per line with at least one coding SNP ranged from 1,025 to 3,446 out of the 4,484 re-sequenced genes, with an average of 2,293 genes per inbred line (S5 Table). In absolute numbers, most coding SNPs were synonymous. In total, 41,334 positions that are polymorphic in at least one inbred line are classified as synonymous, whereas 8,228 are associated with more drastic effects and are annotated as 'non-conservative missense'. The number of nsSNPs in each line ranged from 225 to 2,089, with an average of 1,115 per inbred line. The number of genes with at least one nsSNP ranged from 182 to 1,113, with an average of 625 genes per inbred line. A total of 1,488 genes had a unique nonsense SNP detected in at least one inbred line. In each line, the total number of genes with at least one nonsense SNP ranged from 211 to 590, with an average of 317 genes per inbred line, while the number of nonsense SNPs ranged from 244 to 1,638, with an average of 886 per inbred line. The overall distribution of nonsense SNPs suggests that they are rare variants, with the majority (82.6%) occurring in five or fewer inbred lines.

In total, 45.2% of captured genes were found to harbor at least one nsSNP (S6 Table). Of these, a set of 67 genes was identified as potential candidates for manipulation of biomass production. These were selected by checking the two highest biomass yielders (F2 and F7) and the two lowest biomass yielders (B111 and EA1070; S7 Table) and selecting genes that carry nsSNPs in only one of these two groups (S6 and S8 Tables), either only in the low or only in the high biomass yielders. We observed that 46/67 candidate genes carry nsSNPs only in the high yielding lines, suggesting the modified alleles may be associated with increased biomass. The other 21 genes contained nsSNPs in low yielding lines, perhaps revealing the opposite association. For example, GRMZM2G024374, a 6-phosphofructokinase involved in the glycolytic pathway, contained nsSNPs in the low yielding inbred lines, but none in the high yielding ones. Two genes, AC217050.4, encoding a terpene synthase, and GRMZM2G034069, which is involved in brassinosteroid biosynthesis, contained radical mutations in high yielding lines, but not in low yielding lines.

Presence/absence variation (PAV)

On average, 93 genes (2.2%) per line were classified as completely missing (no read aligned either partially or fully to the target gene sequence), and this ranged from 56 to 169 genes across all studied lines (S9 Table). We further observed that deletion of a given gene relative to the B73 reference was not distributed as simple presence/absence across the studied inbred lines nor was this the case for different genes within an inbred line. If a maximum coverage of 10% of a target gene relative to its sequence in B73 was allowed (90% or more not covered), on average 118 genes per line were classified as missing, ranging from 83 to 215 genes across all studied lines. By relaxing the threshold to a maximum of 20% coverage of a target gene sequence (80% or more not covered with reads), 177 genes were classified as missing, ranging from 114 to 313 across the studied lines. We ultimately declared a gene to be present in a given inbred line if more than 25% of its sequence length (according to the re-sequenced B73) was covered by reads. With this threshold, an average of 220 genes were classified as missing, ranging from 138 to 420 across all studied lines.

Discussion

Maize is a critical crop for human food and livestock feeds; it contains a large and complex genome and exhibits high levels of both phenotypic, and genetic, variation. In this study, we designed a sequence capture assay to identify variable loci within inbred maize lines that may be involved in growth and biomass accumulation. We observed that a significant amount of our 4,648 studied genes accumulated VPs with a substantial proportion of non-synonymous

mutations that might affect the functional integrity of these genes. In addition we observed with a comparative sequence analysis further differences between the studied inbred lines regarding presence and absence of genes that may provide further candidate loci for extended analysis. We are confident that with our in-depth evaluation of computational methods we established a useful approach for diversity studies.

Use of sequence analysis tools

It has previously been observed that variant calling concordance can vary markedly between different tools [17,36,44]. In agreement with these findings, we also observed a substantial incidence of discordant SNP calls. Therefore, we argue that a combinatorial variant calling approach that uses different SNP prediction tools is a prospective practice to achieve results with high confidence. Our evaluation of mapping tools further demonstrated that use of non-optimal read alignment tools might result in a loss of up to 26% of VPs inherent in the analyzed inbred lines. Thus, read alignment has a strong impact on variant calling, and the use of an optimal procedure is of central importance. A great potential exists for re-alignment methods that optimize a constructed alignment with the aim of achieving improved and more reliable SNP calling. Further, recalibration of base quality and optimization of read positioning can also have a large impact on prediction correctness [45,46].

Evaluation of variant calling methods

In this study, we evaluated the merits of various methods for calling high quality variants in 454 sequence data. Based on the final list of variants (383,145 sites; Table 3), the different variant callers varied in performance and displayed comparable patterns of performance values across the four control data sets (50k, GBS, RNA-Seq and HapMap2), in both the overall dataset analysis using standard settings and the results of an in-depth analysis performed on genotype NC358 using variable parameter settings. In regards to the three performance measurements (S_c, S_p, and F_1), the different tools were differentially ranked and displayed comparable ranking patterns when checked with the four control data sets (Table 3). On average, SAMtools (0.94) showed the best sensitivity. Overall, we observed that the average values of these measurements provide a good approximation of global performance. We note that while tools with high numbers of detected VPs have a higher probability of including known true positive VPs, these also predict more false positive sites. However, this does not necessarily reflect the overall exactness of the prediction. To measure this, we compared the specificity (S_p) of the different tools and found that, in this case, the variant callers with low numbers of detected VPs performed best, with SNVer having the best specificity. The F_1-score is an indicator of the optimal combination of sensitivity and precision. Regarding the external control data sets 50k, GBS and RNA-Seq, that have relatively small overlaps with the sequence variants detected in this study, our studied diversity panel assessment revealed that the variant calling methods with lower number of predicted VPs (CRISP, SNVer, VarScan2, and VCMM) showed better performance regarding the F_1-score. Conversely, FaSD, Freebayes, and SAMtools showed better performance for the large HapMap2 control dataset. The sensitivity analysis revealed lower values than one would expect for high correctness. However, it is important to note, that a predicted VP, which is not present in the control data set, is not necessarily a false positive. The corresponding VP may have been excluded from the genotyping array design, or the applied variant calling method was not able to detect that VP (i.e., intronic SNPs are not represented in RNAseq). According to our evaluation of these methods using the sequence data from 21 inbred maize lines, the most reliable polymorphism prediction is achieved when multiple variant callers are cross validated. The resulting combinatorial variant calling approach

performed with higher exactness. With respect to the evaluated eight variant calling methods, we observed best results when a VP was called by at least three methods. Consequently, we are confident that our final diversity set has a higher validity than would be the case if only an individual variant calling method were used.

Large-scale sequence polymorphism discovery

A comprehensive catalogue of genetic variants, including SNPs, INDELs, CNV, PAV, and common and rare variants is an essential resource for studies aimed at identifying those variants that affect phenotypic expression. In this study, we detected a large number of SNP and INDEL sequence polymorphisms. Although they are present at lower rates than SNPs, small INDELs represent a functionally important type of genomic variation [47,48]. In addition we detected numerous cases of PAV among the studied inbreds, which is an important component of genetic variation and hitherto largely untapped. The intraspecific variation of gene content observed even in the limited data set of this study (addressing 4,648 genes in 21 maize inbred lines) indicates that it represents a highly relevant source of genomic variation that can potentially contribute to population's ability to adapt to environmental changes. Therefore, it is important to gain further knowledge of presence absence variation to deepen our understanding of genome diversity and to support the identification of functional variation in traits of interest. We also identified 185,211 novel VPs not present in public datasets, which can be of interest for the design of SNP markers to study variable SNPs with potential functionality or functionally variable haplotypes.

Relatedness of the studied inbred lines and chromosomal distribution of the detected sequence variation

The genetic relationships between 20 inbred maize lines and the B73 reference genome were investigated using the CTD set. A neighbor-joining dendrogram based on SNP profile splits into two main branches, corresponding to the dent and flint groups (Fig 3). These detected groupings were expected and consistent with the historical data on the origin of the inbred lines (S7 Table). For example, UH007 and DK105 are University of Hohenheim breeding materials (flint gene pool). Bouchet et al 2013 working a sample of 375 maize lines representing the worldwide diversity found a similar [49]. The groupings were confirmed by the genetic structuring of the inbred lines, with structure plot giving the position of the inbred lines according to either their known or apparent origin populations (Fig 4 and S7 Table). A majority of the USA lines displayed allelic similarity, with the B series grouping closer to B73 than the others, as would be expected since the name B series (including B73) denotes an apparent origin from Iowa state maize populations (S7 Table). Conversely, the lines originating from France or Germany displayed higher differentiation relative to B73. Overall, the relationships among the inbred lines relative to B73 were reflective of their geographical origin, and thus perhaps a common ancestry, rather than the genetic pool (dent and flint). On the other hand, dents from USA seem to have contributed largely to the genomes of dents in other regions.

Visualisation of SNP density distribution revealed high variation along the ten maize chromosomes. A higher differentiation between the inbred lines (Fig 6) is found at the distal ends, perhaps due to high recombination rates expected in these regions, whereas regions closer to the centromere show lower levels of diversity (S5 Fig). This is in agreement with other observations showing higher conservation at the centromeres [50].

Beyond this overall distribution, several interesting patterns of variation in the density of sequence divergence were observed when the chromosomes were analyzed in bins of 50 kbp size and screened for regions that exceed the 60% quantile of the maximal VP density (across

all inbred lines for particular 50kbp) in at least five genotypes when compared to all other genotypes. Discovered regions are indicated by bars in the second outer circle (S6 Fig). A clear differentiation between dent and flint lines was found on chromosome 3 around the sequence position 109 Mbp. Within this genomic region, a high SNP density was observed in the flints, while a significantly lower SNP density was observed in the dents. The opposite pattern was found on chromosome 8 (114 Mbp), where the dent inbred lines show significantly higher diversity than the flint inbred lines. In addition to the detected loci specific for the dent or flint populations, we found 100 loci of high SNP density. These regions that exceeded the 60% quantile in at least 5 inbred lines sum to ~1.5 Mbp of genic sequences.

SNP annotation

The goal of annotating SNPs is to provide a reference as to which ones may be functionally relevant. Our detected SNPs were assigned to a diverse range of functional classes, with the majority classified as silent mutations. These may have no or little effect on the phenotype. Among all classes of SNPs detected, nsSNPS are the most likely candidates for causal mutations, as they could alter the structure and function of relevant proteins. Such genetic variation may thus account for substantial trait variation in the studied lines. We observed an average of 984 nsSNPs per line in gene coding regions (Table 4), with a total union of 8,228 nsSNPs across all inbred lines. In humans, nsSNPs in gene coding regions could account for nearly 50% of the known genetic variations linked to human inherited diseases [51]. Thus, a larger effort would be warranted to study potential links between the identified nsSNPs and trait variation in maize, and to determine and how they affect the regulation of biological pathways and processes. The 1,170 SNPs that create premature stop codons and the 4,131 SNPs that are predicted to affect splicing can also be predicted to have particularly pronounced effects (Table 4). SNPs selected or prioritized in this way would be highly preferred marker sets to be subjected to association studies using suitable larger populations such as the CornFed Dent and Flint panels for which very substantial genotype and phenotype information has already been collected [52,53].

Gene diversity and functional annotation of target genes accumulating radical mutations

The majority of genes displayed low sequence polymorphism across all studied lines, with an average of 10 VPs per target gene. This translates to a global average SNP density of 1/639 bp within these target genes, suggesting that the majority may be involved in key plant pathways and hence contain a low number nucleotide changes. This correlation between low gene diversity and its importance for the organism has been emphasized in previous studies [54]. Relating SNP densities in exons and in introns, we observed a ratio close to 1 (Table 4). The nucleotide variability in the non-coding part of a gene is expected to be higher than the variability observed in the coding part, because non-coding polymorphic sites are less likely to cause structural changes of encoded proteins that could lead to functional consequences [55]. However, because of the capture array design, which required exclusion of repetitive sequences, we postulate that this ratio is slightly skewed towards an underestimation of polymorphic sites in introns. Previous studies have demonstrated the insertion of repetitive elements into the non-coding sequence of maize genes [56], and consequently, repetitive probes that originate from intron sequences would have been discarded in our scheme, resulting in their consistently lower representation on the array. In order to test this hypothesis, we analyzed the sequence coverage within exon and intron sequences. In all studied maize inbred lines, the calculated coverage in coding regions exceeds the coverage in non-coding gene sequences (on average

by ~15.2%). Beside the exclusion of repeats in the design of the array an excess of coverage in coding sequences can also be explained by a higher degree of sequence conservation as compared to the non-coding sequences. High sequence similarity is a requirement for successful sequence capture that would tolerate only a minute level of nucleotide variation. These results highlight the fact that the properties of the capturing system needs to be considered for correct interpretation of results, especially if different levels of sequence conservation can occur and they indicate a better performance of the approach in the analysis of coding versus non-coding sequences.

A much more detailed perspective can be considered on the gene level when utilizing the complete diversity files (S1 File) or the list of genes with radical mutation (S6 Table). These genome regions, which include 344 of the target genes, can be regarded as candidate loci for further studies aiming at linking genetic variation to phenotypic variation related to yield formation / biomass production in maize. A further refinement in the nomination of candidate genes is possible within the existing data set, when the differential occurrence of radical mutations in high and low biomass producing lines is taken as selection criterion. Although, due to the very restricted number of lines taken into account in the example given in this study (2 low and 2 high yielding lines), no statistical support can be given to the relevance of the selected 67 genes (S8 Table), they could be regarded as prime candidates for targeted association testing. This is supported by the suggested roles of some of the encoded gene products in plant growth:

The GRMZM2G034069 gene is involved in brassinosteroid biosynthesis and showed an enrichment of radical mutations in high yielding lines, as compared to low yielding ones. Brassinosteroids are a major hormone class controlling plant growth and development [57]. In contrast, GRMZM2G110881 displayed an accumulation of radical mutations in low yielding lines and not in high yielding ones. It encodes a UDP-glucose 4-epimerase, which converts UDP-galactose to UDP-glucose and is involved in glycosyltransferase reactions in metabolism. GRMZM2G393762, which encodes a pectinesterase, also showed an accumulation of radical mutations in low yielding lines, as compared to high yielding ones. Pectinesterases are involved in plant cell wall modification and subsequent breakdown. Additionally, many genes involved in plant defense reactions (e.g., AC211164.5 and GRMZM2G702176) have accumulated radical mutations in the low yielding lines but none in the high yielding ones.

Presence absence variation

Another level of genome diversity / sequence variation, which is increasingly recognized and considered as a further potentially important determinant of trait variation in maize [2–4] is presence / absence variation. Numerous gene sequences of up to several kilobases in length were found in this study to be absent in at least one of the studied inbred lines. These sequences may be common in the maize population, and thus their absence in the genome of a given line might be a rare variant. Using a threshold cut-off of at least 75% gene sequence loss relative to the reference gene sequence in the B73 genome, 2.2% of the targeted genes were found to be absent in at least one of the studied lines.

When the two lowest and two highest biomass yielders were compared, 30 presence / absence genes of interest were identified (S10 Table). Among them are three genes (GRMZM 2G048775, GRMZM2G050829, and GRMZM2G129935), which encode peroxidases and are possibly involved in the betanidin degradation pathway. GRMZM2G048775 and GRMZM2 G050829 are present in the genomes of low yielding lines but absent from the genome of high yielding lines, while GRMZM2G129935 is only present in high yielding lines. Plant peroxidases are encoded by a large multigene family, and are known to participate in a broad range of physiological processes, such as lignin and suberin formation, cross-linking of cell wall components,

and synthesis of phytoalexins [58]. They are also known to participate in the metabolism of reactive oxygen species and reactive nitrogen species, both of which trigger the hypersensitive response, a form of programmed host cell death at the infection site, associated with limited pathogen development [58]. Most genes in this category were found in the genomes of low yielding inbred lines and absent from genomes of high yielding lines. Other identified candidate presence/absence genes include ones involved in glycerol degradation (e.g., GRMZM2G 079100), plant sterol biosynthesis (e.g., GRMZM2G014789), glycogen biosynthesis (e.g., GRM ZM2G481027, also called glycogenin glucosyltransferase), acyl carrier protein metabolism (e.g., GRMZM2G4181199), and adenosine nucleotides *de novo* biosynthesis (e.g., GRMZM2G 470035) pathways. Additionally, an enzyme encoded by GRMZM2G079100 (sn-glycerol 3-phosphate:ubiquinone-8 oxidoreductase) is essential for post-germination growth and seedling establishment [59]; this gene was found to be present in high yielding inbred lines but absent from low yielding lines, possibly indicating the importance of seedling establishment to the final biomass. In addition to the genes showing differential occurrence of radical mutations in high and low biomass producing lines, the differentially present / absent genes share the same level of support and should thus as well be considered with priority in follow-up experiments towards testing their associations with high or low biomass accumulation.

In summary, the targeted re-sequencing of a large list of selected genes across a series of diverse maize inbred lines differing in biomass accumulation revealed a high degree of DNA sequence polymorphism. We were able to narrow down several candidate loci that may play a role in the phenotypic differences between different maize lines, and these results may be considered for future studies aimed at optimizing yield in this important agricultural resource.

Materials and Methods

Materials

Maize lines were obtained from the PLANT-KBBE-CornFed project: B101, B102, B106, B111, B73, F353, F7059, Mo17, Mo24W, NC358, P068, P107, P128, and PH207 are from the CornFed Dent panel, and DK105, EA1070, EP1, F2, F7, Lo32, and UH007 are from the CornFed Flint panel [52,53]. These inbred lines were chosen to represent a wide range of biomass production properties. The geographic origin and a short description of the maize inbred lines used in this study are presented in S7 Table.

Methods

Sequence capture array design. The sequence capture array was designed to target the full-length enrichment of gene sequences and was geared for the generation of re-sequencing data for complete gene haplotypes. An inventory of 4,823 candidate genes representing possible targets for modification of biomass accumulation and production, as well as water use efficiency, was compiled from the published literature (S11 Table).

A BLAST of the candidate genes from different species (maize, rice, barley, and *Arabidopsis*) was carried out against maize CDS (maize genome project: www.maizesequence.org, version ZmB73_4a.53). Maize gene identifiers were checked for uniqueness, and genomic sequences (exons, including introns) were extracted. All genes were extended by 1 kb at the 5′ and 3′ ends, and the primary target regions to be placed on the maize array were determined. This yielded a set of sequences having an I.50 of 6,583 bp, with the largest contig covering 77 kb. Using this information, and eliminating redundancy, the number of genes was reduced to 4,758 (S12 Table), corresponding to 99% of the 2.1 M NimbleGen microarray capacity. These genes were well distributed across all the maize chromosomes (Fig 7). The gene coordinates

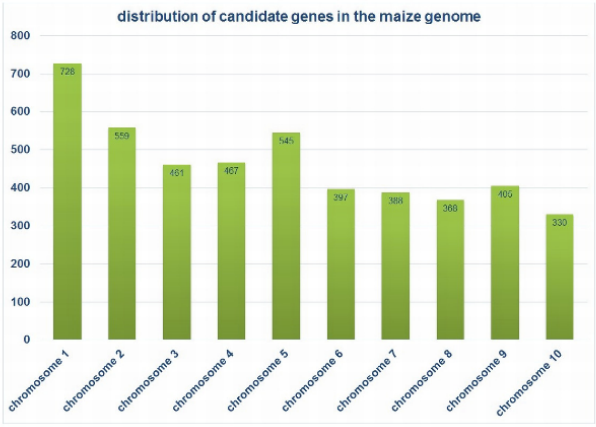

Fig 7. Distribution of 4,785 candidate genes over 10 maize chromosomes. Selected genes are predicted to control various aspects of plant growth, biomass production, and composition; bar graph depicts their distribution in the maize genome.

doi:10.1371/journal.pone.0132120.g007

genes were then sent to Roche NimbleGen to facilitate the design of a custom 2.1 NimbleGen sequence capture microarray.

NimbleGen array design and algorithmic details for probe design. To avoid non-specific binding, repetitive elements were excluded from the probe design using the RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) and WindowMasker [60] software packages. A frequency histogram of all 15mers in the target genome was constructed, and the frequencies of all 15mers forming a probe were analyzed. The average 15mer frequency was computed, and only probes having less than a 100-fold frequency were used. Uniqueness of the probes was determined by the Sequence Search and Alignment by Hashing Algorithm (SSAHA) program [61] developed at the SANGER Institute (http://www.sanger.ac.uk/resources/software/ssaha/). SSAHA probes were compared to the reference genome from which they were generated (Maize genome AGP v1, release 4a53; www.maizesequence.org). For each oligonucleotide, mapping determined if the probe matched either perfectly or closely to the genome, using a word size of 12 bp. A heuristic then accepted an oligonucleotide if its minimal match size was above the oligonucleotide length minus the word length. Therefore, if we assume an average oligonucleotide size of 50 bases, a minimal match would be expected to be at least 38 bp in length. In our design, a more stringent setting was used, allowing a minimum match size of 30. The mismatches were weighted in a trapezoidal shape, giving a higher weight to a failure in the middle of the oligonucleotide than to its end. If, in any comparison to the genome, the mismatch score was less than 10, the probe was rejected as non-unique. To use the full sequence capture array capacity, in the final probe selection replicated probes were assigned.

Evaluation of uniqueness settings. The design of the sequence capture array was balanced between (1) unique probes (MM1), resulting in a less optimal genome representation due to the masking of paralogous and conserved sequence domain of gene families, and (2) multiple matching probes (MMX), arrived at by the use of a less stringent probe uniqueness setting and resulting in a better coverage of the selected target genes. NimbleGen provided five independent designs with various stringency settings (MM1, MM2, MM4, MM5, and MM10). After a detailed investigation of the proposed probe sequences, we opted for MM5 (S13 Table) in order to avoid masking to many conserved gene families and/or paralogs. The consequence of

lowering the uniqueness criteria is that more genes are allowed to match a given oligonucleotide, and thus specificity is reduced. However, our analysis revealed that MM5 has the optimal balance between specificity and required coverage. Using MM5, the designed Sequence Capture array covered 53.4% of target bases and 94% of our target genes, with an average probe length of 75 bp (S14 Table). This array may be ordered from Roche NimbleGen by requesting the design: 110308_ZmB73_AGP_v1_MM_cap_HX1.

A total of 4,648 target genes (97.7% of the initially targeted genes) were included in the final sequence capture array. NimbleGen utilized 2.1 million oligonucleotide probes, and if adjacent probes were in close proximity to one another they were compiled as a 'probe cluster'. On average, each gene had 60% of its sequence covered by oligonucleotides through the use of 23,252 probe clusters.

The AGPv1 probes for the sequence capture were anchored to the maize reference (v3). The nature of the maize genome is complex, with many nearly identical paralogs. We tried to circumvent this complexity in the design of the sequence capture array via the balanced use of the multiple match parameter 'MM5'. With our setting we successfully could limit the proportion of paralogous probes. We only observed a proportion of 2.3% of the 23,252 designed probe clustering that had an identical paralogous sequences. For further coverage analysis and estimation of the number of gaps between probes in the target region, only the uniquely anchored probes were used. Generally, probe coverage was higher in exon regions (69.2%), as compared to introns (27.0%) and the UTR regions (14.3%; S18 Table). Lower probe coverage can lead to an increase in reference positions that are not captured, literally referred to here as 'gaps'.

We analyzed gaps between adjacent oligonucleotides in order to determine the number and sizes of gaps occurring in the target regions. We observed a well-optimized representation of target genes, especially for exons (S18 Table). Exon regions displayed fewer and shorter gaps, with only 1,801 gaps over 300 bp and 551 gaps of more than 500 bp in size. However, intron and UTR regions were characterized by a higher number of longer gaps, with more than 4,500 and 2,500 gaps in intron and UTR regions, respectively, of more than 300 bp in size (S18 Table). Our metrics show that 45% of the 26,809 total gaps are longer than 300 bp, and these large gaps are less likely to be spanned by short sequence reads. Moreover, it was expected that the number of gaps might increase when DNA from highly diverse maize lines is captured. Based on this analysis, we decided to use 454 sequencing technology, which is characterized by long sequence reads. In addition, the longer 454 reads are beneficial for correctly aligning them to the reference sequence, particularly across the highly conserved stretches of closely related genes. Consequently, 454 reads are advantageous to avoid ambiguities in the alignment of repetitive and paralogous sequence stretches (including sequence variation within these stretches).

Genomic DNA extraction. Fifteen seeds from each of the 21 inbred lines were planted in pots containing commercial potting soil mixtures in a greenhouse at the Institute of Plant Genetic and Crop Plant Research (IPK-Gatersleben), Gatersleben, Germany. The greenhouse has supplemental illumination using SonT Agro high pressure sodium lamp (Philips, Amsterdam, Netherlands) set to 16 h of light per day. Genomic DNA (gDNA) was isolated separately from 5 cm leaf samples of 10 randomly selected 2-week old maize seedlings from each inbred line using a modified CTAB protocol [62]. DNA was then purified by proteinase K digestion, followed by ammonium sulfate precipitation. DNA concentration and quality were assessed using a Nanodrop spectrophotometer (Willmington, DE) and electrophoresis in 0.7% (w/v) agarose gels, to verify integrity. After the normalization of DNA concentration to 250 ng/μl, equal amounts were pooled from the 10 individuals per genotype to constitute the working gDNA.

DNA capture. For DNA capture, 500 ng of each DNA sample was nebulized to yield fragments of approximately 250 bp to 1 kb in size. In brief, the fragmented genomic DNA samples were polished to form blunt-ended fragments, adaptors were ligated onto these fragments, and small fragments were removed. The libraries were quantified using a fluorometer, and quality was assessed using an Agilent Bioanalyzer high sensitivity DNA chip (Agilent Technologies, Santa Clara, CA, USA) to ensure libraries had a mean fragment length between 600–900 bp, with a lower size cut-off less than 10% below 350 bp, as recommended by NimbleGen protocols. Libraries that satisfied these characteristics were amplified by ligation mediated (LM)-PCR. Each sample was then evaluated with an Agilent Bioanalyzer 7500 (Agilent Technologies) to ensure a mean fragment length between 600–900 bp, as per the NimbleGen Sequence Capture protocol. The amplified sample libraries were then checked for quality to ensure the A_{260}/A_{280} ratio ranged between 1.7–2.0 and sample library yield was $\geq$ 1.5 μg, as per NimbleGen recommendations. Subsequently, 30 μl of Plant Capture Enhancer (PCE, Roche NimbleGen, Basel, Switzerland) was added to 1.5 μg of each sample library that satisfied these characteristics, herein referred to as pre-capture amplified library, to block repetitive sequences. Hybridization buffer and hybridization component A (Roche NimbleGen) were added to each sample and loaded to a 2.1 M NimbleGen sequence capture array. The hybridization was done for 72 h at 42°C.

On completion of hybridization, each slide was washed according to NimbleGen protocols, and the bound library was eluted with 50 μl of nuclease free water. The eluted DNA was amplified by LM-PCR, and the resultant libraries, herein referred to as captured DNA, were characterized to determine concentration, size distribution, and quality. The captured DNA samples showed fragment distribution ranges from 500–1300 bp, with distribution peaks between 650 and 1000 bp.

Following the completion of the amplification reaction, samples were purified using a Qiagen QIAquick column (Qiagen, Venlo, Netherlands) following the manufacturer's recommended protocol. The DNA was then quantified using a NanoDrop-1000 (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and electrophoretically evaluated with an Agilent Bioanalyzer 7500 chip (Agilent Technologies). Aliquots of the resulting captured DNA sequencing libraries were diluted to 2 x 10^5 molecules/μl for GS FLX sequencing.

Determination of captured genomic DNA enrichment. A total of 52 loci, representative of the 10 maize chromosomes, were selected from among the regions targeted by unique probes, and qPCR primer sets were designed. These were all tested together, along with one locus provided by NimbleGen (NSC-247), using the genomic DNA from different maize genotypes. Four loci that displayed good amplification across the tested genotypes were finally selected (S15 Table and S7 Fig); these are located on chromosomes 1, 3, 6, and 8.

To measure the enrichment of the captured DNA, and thus the capture efficiency, quantitative fluorescence PCR was performed on pre-captured and post-captured enriched libraries using an ABI RT PCR system. For each locus, fold enrichment = $(E)^{\Delta\text{-Ct}}$ was calculated, where E is the efficiency of the amplification, and Ct is the point at which the generated florescence signal rises above that of background. *E* was generated empirically for each locus using LinReg PCR analysis of Real Time PCR, and Δ-Ct was calculated by subtracting the Ct of the captured library from that of the non-captured library.

GS FLX DNA sequencing and raw data analysis. Captured DNA was sequenced with the GS FLX instrument at the IPK-Gatersleben (Genome Centre). The Low Molecular Weight DNA Protocol was used to prepare the 454 GS FLX sequence-ready libraries. DNA sequencing libraries for all 21 samples were prepared separately before the amplification by emPCR, following the steps described in the GS FLX emPCR Method Manual. After emPCR amplification, two prepared samples were each loaded in a half gasket PicoTiterPlate device (70_75 mm;

Roche/454) and sequenced in a GS FLX system with standard Roche/454 protocols. The 454 pyrosequencing data were collected after a 7 h run on the GS FLX system, and the Roche/454 gsMapper (454 Life Sciences) was used initially to analyze all raw sequence reads that were generated.

Sequence analysis and alignment. The 454 reads were trimmed and filtered as described previously [29] to ensure the high data quality for all downstream analyses. To evaluate the raw read quality, clc_quality_trim (CLC assembly cell, version 4.2) was used to trim reads that did not achieve sufficient base pair quality standards (minimal quality PHRED score >20), excluding short sequences (<30 bp).

Maize genome references. Sequences were aligned to the B73 reference genome (AGPv3.20) using various read alignment programs. The genomic reference and corresponding gene models (GFF, cDNA, and protein sequence) were downloaded from Gramene [63] (http://ensembl.gramene.org/Zea_mays/Info/Index).

Conversion of target regions to maize reference v3 (AGPv3.20). The design of the capture array was based on the maize reference AGPv1. However, all mappings and further downstream analysis were performed on the AGPv3.20 reference. To anchor AGPv1 based MM5 positions to the AGPv3.20 maize reference, two approaches were applied. For target genes with an identifier present in both reference versions, their corresponding gene positions were used. For the 306 genes with no corresponding gene identifier in v3 reference, these target regions were anchored to the AGPv3.20 reference by performing BLASTN [64] analyses with an identity threshold of 98%, using BLAST version 2.2.28. Using these two approaches, all MM5 target gene positions were anchored to AGv3.20 reference (S13 Table).

***De novo* assembly.** A *de novo* assembly of the quality trimmed 454 reads was performed using Newbler (version 2.6, 454 Life Sciences) with the default settings. We also re-sequenced B73 for our targeted regions with the 454 platform and mapped it to the respective maize reference genome. This was used as a control to identify regions of the B73 assembly that were captured by our designed array. The resulting genotype specific assemblies were analyzed with BLASTN (identity 99.8) against the selected target gene reference models (cDNA) to estimate the presence of a gene. A threshold of 75% gene sequence length absence relative to the parent gene sequence length in reference was used to as model to define absence of a gene in a given inbred line.

Read alignment evaluation. Seven read alignment tools (BWA-SW, BWA-MEM, Bowtie2, CLC mapper, Smalt, Stampy, and NextGenMap) were used to evaluate read alignment methods. This collection of alignment tools consists of widely used standard tools, as well as recently published tools that are able to handle Illumina and 454 reads. For BWA, the previous version, BWA-SW, for long reads, and a recently released version, BWA-MEM, which also is suitable for long reads, were used. Stampy is the preferred read aligner for Illumina reads but performed well in our study in handling 454 reads. CLC mapper is a commercial tool released within the CLC assembly cell repository. An in-depth evaluation of read alignment tools was performed on a large-scale utilizing all 21 genotype lines, in three independent parameter configurations. Besides the standard parameter settings, we utilized for each read aligner, two additional custom settings aiming for a higher sensitivity. The detailed settings are provided in S2 File. To compare mapping results, the GATK [46] tool repository was used. Utilizing the combined reads of all 21 inbred lines, regions with sufficient coverage (—cov 20) were screened and then extracted with 'FindCoveredIntervals'. Coverage in target regions was discovered with 'DepthOfCoverage' using our target coordinates (S4 Table). Genes absent in certain inbred lines were detected by applying BEDtools 'coverage' software [65] to the individual mappings (BWA-MEM) per inbred line.

Variant detection. A wide range of read alignment and SNP calling tools was used to establish the most favorable set of SNPs in our maize panel, and VCFtools [66] was utilized to calculate statistics for each SNP data set. A high number of INDELs is expected, since 454 reads have an inherently higher error rate at long homopolymers stretches [67], which can lead to false positive INDEL detection. To avoid this, the maize genome was screened for homopolymers equal to, or larger than, 7 bp, and the detected INDELs that were embedded, or adjacent to, homopolymers were classified as false positives and discarded. Additional criteria, such as that a basic INDEL position must be supported by at least two variant callers, were applied to further filter the INDEL positions. Two approaches can be utilized to filter VPs: the first is an internal filtering, with parameter settings that aim to reveal a more stringent variant calling. In addition, subsequent to the variant calling process, predicted candidates can be analyzed and filtered. This *posterior* filtering plays an important role in variant discovery and has a crucial impact on reaching a high quality prediction [68]. However, in our study we tried to estimate the sensitivity of a variant calling method without the influence of a user defined parameter setting and therefore compared the unbiased default settings. In addition, an in-depth evaluation of variant calling methods was performed, utilizing a randomly selected inbred line ('NC358'). For this selected line, we utilized all 21 pre-calculated alignments for subsequent variant calling. Besides the standard parameter settings, we used two additional custom settings for each variant calling method, aiming for a higher sensitivity, and totaling to 504 tested combinations. The detailed settings are provided in S2 File.

Diversity assessment. To measure the detection power and correctness of each evaluated variant caller, we calculated sensitivity, specificity, and F_1-score, as shown below:

$$Sensitivity\ Se = \frac{number\ of\ true\ positive\ sites}{condition\ positive\ sites} = \frac{TP}{(TP + FN)}$$

$$Specificity\ Sp = \frac{number\ of\ true\ negative\ sites}{condition\ not\ polymorph\ sites} = \frac{TN}{(TN + FP)}$$

$$F_1 score \quad F_1 = \frac{(2 * TP)}{((2 * TP) + FN + FP)}$$

The sensitivity measure indicates the power of a tool to detect true positive sites (designated TP in the equations above). The specificity estimates the true negative (designated TN in the above equations) detection rate, indicating the ability of a particular variant calling tool to discard true negative calls. The F_1-score is the harmonic mean, balancing precision and sensitivity. Four control data sets were investigated to evaluate and validate variant predictions: 1) the genotyping array (50k) comprising of 52,374 SNP markers [5], 2) the genotyping-by-sequencing (GBS) comprising of 719,487 [39], 3) the RNA-Sequencing (RNAseq) comprising of 931,484 markers within the gene space of maize [40], and 4) the large data set of the maize HapMap2 (HapMap2) comprising of over 55 million SNPs (http://data.iplantcollaborative.org/quickshare/e75bc315fc0f9fda/HapMapV2RefgenV220120328.vcf.gz) [41].

A *'true positive site'* refers to a variant position that was predicted in our data set and is matching a position in the corresponding control data set. The intersection of a particular control dataset and the detected VPs that are validated by at least one of the evaluated variant calling methods or settings defines the *'condition positive sites'* (CP) represented in our data (Table 3). Subsequently, these CP are used to classify a prediction of a particular variant calling method as *'true positive'* or *'false positive'* (FP). As described above, the F_1-score is defined as the harmonic mean between sensitivity and precision, where the latter estimates the ratio

between TP and FP. For the F_1-score calculation, we extended this set to include all VPs that have an overlap to our final 'CornFed Target Diversity' (CTD) dataset. This strategy aims to circumvent the severe underestimation of true positive sites that would occur if only the less pronounced overlap to the control dataset was classified as TP, where all other predictions would be automatically discarded and judged as FP. Finally, the set of *'false positive'* sites is defined as the opposite of the TP, meaning that a FP site is not in the control dataset and not in our CTD set. FP sites refer to VP candidates that do not have enough support by independent variant calling, as defined by the CTD setting (three independent variant calling methods). We then define all positions that have on average a minimal coverage of one read, but were not called as polymorphs, as *'true negative'* (TN) sites. In conjunction with sites classified as FP, these TN sites define the complete set of '*condition not polymorph sites' (CNP)*. To evaluate the prediction power and the optimal detection method, we assessed each of the eight variant calling tools and analyzed their predictions for each of the 21 maize inbred lines. This evaluation has been extended for one genotype (NC358) by using multiple parameter settings (S2 File). To assess the overlapping and non-overlapping VPs among the four control and CTD data sets we constructed a diagram for illustration. The Venn diagram was created using the web tool provided by the Bioinformatics and Systems Biology of Gent (http://bioinformatics.psb.ugent.be/webtools/Venn).

Genotype concordance. All the 21 maize inbred lines re-sequenced at specific target genes were also genotyped with the 50k maize genotyping array [5] (S2 Table). Individual variant discovery was also performed for each genotype with the eight variant callers. Subsequently, considering our target genomic regions that were also genotyped with 50k array, the overlap between the 50k genotyping and that of the variant callers was computed, and the called alleles were analyzed for genotype concordance. Eighteen of the 21 inbred lines (excluding Lo11 and Mo24W in which the analysis failed, and B73 which is the reference genome) were used in this analysis.

Combinatorial variant calling approach to detect final SNPs. The final SNP set was established by posterior filtering. Respective positions fulfilled the following criteria: 1) VP has a minimal read coverage of five, 2) VP is in bi-allelic sites, 3) VP exceeds the minimal quality score (>0.4 normalized quality score), and 4) VP is predicted with at least three independent variant calling methods. This cross validation parameter of three tools was investigated to verify a prediction and to achieve an optimal result within our heterogeneous maize collection. The threshold of three independent tools supporting a prediction was defined to be the optimal trade-off between sensitivity and specificity. Therefore we validated various thresholds (1–8) and their combined prediction power using the results of the 50k genotyping data set of all 21 inbred lines. To calculate the F_1-score false negative and false positive sites had to be considered. FN sites are VPs that were missed from the 50k array (proportion with overlap to our target sequencing) and missed from the CTD set. FP sites are positions that were predicted but have no overlap to any of these two control sets.

SNP annotation. The final SNP set was annotated using the software COOVAR [69].

Genetic relationship between 21 inbred maize lines. Phylogenetic clustering of the final SNP data set was performed with SNPhylo [70] with standard settings. A graphical representation was generated with Circos [71] using the final set of SNPs. A total of approximately 1 million variants detected in all 21 inbred lines were used to generate a graphic presentation. For clarity of presentation, SNP densities were displayed in regions of 500 kbp (window size).

Population structure. The software STRUCTURE [72,73] was used to analyze the population structure based on the admixture model, where each individual draws some fraction of its genome from each of the K populations. This method is useful to identify population stratification, since inbred lines whose genotypes indicate admixture are assigned jointly to two or more

populations. The correlated allele frequencies model [74] which often improves clustering for closely related populations was used. Ten runs of STRUCTURE were carried out for each set of K sub-populations, with K values from 2 to 10. The choice of the number of K (2–10) was based on earlier cluster analysis using SNP profile, which gave an estimate of 7–8 grouping of the studied inbred lines. The ad hoc [75] criterion was used to determine the optimum value of K.

Presence and absence of genes. To assess the absence of genes in the studied inbred lines, we used BEDtools [65]. The coverage of captured genes was computed from all BWA-MEM generated read alignments (BAM files). To discover the sequence captured for each gene, we first analyzed the presence of target gene reads in the re-sequenced B73 454 sequence reads. Subsequently, the presence of the represented proportion was analyzed in all other re-sequenced maize inbred lines. Minimal read depth threshold was set to 2, and the covered gene length was analyzed with different thresholds (minimal gene sequence length 0%, 10%, 20%, and 25%) to determine if a given gene is present in, or absent from, a given inbred line. A gene was declared present in any given inbred line if more than 25% of its sequence length (according to the re-sequenced B73) was covered by reads; otherwise, it was declared absent. Complete target gene sequence length coverage in the B73 reference was determined using two thresholds, i.e., 80% and 90%. To further validate the absence of a gene, a BLASTN [64] analysis was performed on the *de novo* assembly of a given genotype for a given gene declared as absent.

Functional annotation of target genes. Functional annotation of the target genes was carried out using BLAST2GO [76]. The analysis of the three ontology classes (biological processes, molecular function, and cellular component) was performed for all the target genes. The third level of the GO hierarchy was used to subdivide the gene set into clusters (S16 Table and S8 Fig). We then analyzed the resulting categorization for presence-absence genes and for presence of radical mutation on the target genes. In addition, for the final candidate genes, pathway and a further enzymatic annotation from MaizeCyc [77] information were integrated.

Accession numbers. The raw 454 DNA sequence data obtained from sequence capture and re-sequencing of the 21 maize inbred lines have been deposited at the European Nucleotide Archive (ENA) in the Sequence Read Archive (SRA) and are available under the following EBI project ID: PRJEB5496, [http://www.ebi.ac.uk/ena/data/view/PRJEB5496]. Details are provided in S17 Table.

Supporting Information

S1 Fig. In-depth evaluation of read alignment methods, utilizing three different parameter settings. For each particular setting, the mapping is evaluated for all 21 inbred lines by determining the number of aligned reads, the number of aligned read on target, the mutual agreement of the alignment ('number of reads mapped by other method'), the uniquely aligned reads, and the number of reads aligned by the majority of methods. The graphic shows the results for genotype NC358, which was selected as example.
(TIF)

S2 Fig. Heat map showing the impact of read alignment (seven read alignment methods) on diversity detection (eight variant detection methods). The analysis was performed on the random selected genotype 'NC358', and all 504 possible combinations of alignment and variant calling methods (three parameter settings) were included. To construct the heat map, the detected VPs of each individual approach were compared to the 50k data to reveal true positive predictions.
(TIF)

S3 Fig. Application of different cut-off values for the variant caller count (VCC). All VPs in our studied lines that overlap a 50k position are considered true positive. The proportion of the total number of true positives is depicted in blue for each VCC value. The F_1-score, shown in yellow, illustrates the impact of false positive and false negative values. The harmonic mean reaches highest values at VCC3.
(TIF)

S4 Fig. STRUCTURE analysis to estimate the value of K for optimal partitioning of data.
(TIF)

S5 Fig. Genotype independent distribution of diversity density, shown for the 10 maize chromosomes. Chromosome bins with a size of 2 Mbp were analyzed, and regions that are characterized with high diversity in our genotype collection are indicated in yellow.
(TIF)

S6 Fig. Circos histogram plot for the captured target gene diversity in 21 maize inbred lines. Chromosomes were analyzed in 500 kbp bins, and the genotypes are shown in the identical order as depicted in Fig 6.
(TIF)

S7 Fig. DNA gel electrophoresis for qPCR control loci.
(JPG)

S8 Fig. Pie chart of Blast2GO annotations of biological processes, molecular function, and cellular components within the set of target genes. The Blast2GO hierarchy is presented at level three for all three categories.
(TIF)

S1 Table. Measurement of B73 sequence enrichment using qPCR.
(XLSX)

S2 Table. Genotyping of 21 maize inbred lines using 50k Illumina array.
(XLSX)

S3 Table. Statistics of *de novo* assembly for 21 maize inbred lines.
(XLSX)

S4 Table. SNP annotation for all 21 maize inbred lines.
(XLSX)

S5 Table. Survey of the SNPs in coding regions of the re-sequenced genes in each of the inbred lines. The number of non-synonymous coding SNPs and nonsense SNPs, and the number of genes they affect is also shown.
(XLSX)

S6 Table. List of genes that are affected by a SNP mutation leading to radical protein changes for each genotype.
(XLSX)

S7 Table. Geographic origin, including pedigree data, of genotypes, and a short description of maize inbred lines.
(XLS)

S8 Table. Joint analysis of phenotypic information (high and low biomass) in combination with existence of a radical mutation identified in candidate genes.
(XLSX)

S9 Table. List of captured target genes providing a comprehensive overview of the capture array sequencing. The resulting sequence coverage is presented as percentage of gene length.
(XLSX)

S10 Table. List of captured target genes, including the assigned biological function. A joint analysis links the phenotypic information of lowest yielding (B111 and EA1070) and highest yielding (F2 and F7) inbred lines with the presence and absence information. A gene was declared absent if at least 75% of sequence length is missing, and the captured gene length of maize inbred B73 is used as reference.
(XLSX)

S11 Table. Inventory of candidate genes for the NimbleGenArray design, including a description of target genes predicted to be involved in the modification of biomass accumulation and production, as well as in water use.
(XLS)

S12 Table. Candidate genes selected for designing the 2.1 M NimbleGen sequence capture microarray.
(XLS)

S13 Table. NimbleGen array design providing positions of selected target genes anchored in AGPv1 and AGPv3.
(XLSX)

S14 Table. Array design statistics.
(XLS)

S15 Table. Forward and reverse primer sequences for four different IPK control loci.
(XLSX)

S16 Table. Blast2GO annotation for biological processes, molecular function, and cellular components identified within the set of target genes.
(XLSX)

S17 Table. EBI submission details of the raw data.
(XLSX)

S18 Table. Extended analysis of the NimbleGenArray design. The analysis included 4,342 non-overlapping genes that were mapped uniquely to the maize genome reference (v3). Results revealed different representations of sequence capture probes in UTR, exonic and intronic regions of studied genes.
(XLSX)

S1 File. List of 383,145 variant positions, including SNP annotation and PIC information.
(TXT)

S2 File. List of parameter settings that have been applied for each of the integrated read alignment and variant calling method.
(TXT)

Acknowledgments

The authors acknowledge the technical suppot of Beatrice Knüpffer and Susanne König (Leibniz Institute of Plant Genetics and Crop Plant Research) for their help with DNA sequence capture and sequencing, respectively. Furthermore, we thank Alain Charcosset and Cyril

Bauland (INRA, Moulon) for the provision of seed material of the studied maize lines, and we are grateful to all partners of the CornFed project consortium who gave input on defining the target gene list. We acknowledge Eva Graner and Martin Ganal (TraitGenetics GmbH) for assistance and data access to the maize 50k SNP genotyping array. We would also like to acknowledge Doreen Stengel for the submission of sequence data.

Author Contributions

Conceived and designed the experiments: TA US TS MMM. Performed the experiments: MMM TS. Analyzed the data: TS MMM CU US TA. Contributed reagents/materials/analysis tools: TS MMM. Wrote the paper: TS MMM US TA. Developed the bioinformatics' approaches: TS.

References

1. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; 326: 1112–1115. doi: 10.1126/science.1178534 PMID: 19965430
2. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, et al. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 2009; 5: e1000734. doi: 10.1371/journal.pgen.1000734 PMID: 19956538
3. Swanson-Wagner R a, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 2010; 20: 1689–99. doi: 10.1101/gr.109165.110 PMID: 21036921
4. Beló A, Beatty MK, Hondred D, Fengler K a, Li B, Rafalski A. Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet. 2010; 120: 355–67. doi: 10.1007/s00122-009-1128-9 PMID: 19756477
5. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, et al. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One. 2011; 6: e28334. doi: 10.1371/journal.pone.0028334 PMID: 22174790
6. Riedelsheimer C, Lisec J, Czedik-eysenberg A, Sulpice R, Flis A, Grieder C. Genome-wide association mapping of leaf metabolic pro fi les for dissecting complex traits in maize. 2012; doi: 10.1073/pnas.1120813109/-/DCSupplemental www.pnas.org/cgi/doi/10.1073/pnas.1120813109
7. Andersen EC, Gerke JP, Shapiro J a, Crissman JR, Ghosh R, Bloom JS, et al. Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. Nat Genet. Nature Publishing Group; 2012; 44: 285–90. doi: 10.1038/ng.1050
8. Buckler ES, Gaut BS, McMullen MD. Molecular and functional diversity of maize. Curr Opin Plant Biol. 2006; 9: 172–6. doi: 10.1016/j.pbi.2006.01.013 PMID: 16459128
9. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell. 2014; 26: 121–35. doi: 10.1105/tpc.113.119982 PMID: 24488960
10. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012; 44: 217–20. doi: 10.1038/ng.1033 PMID: 22246502
11. Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, et al. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol. 2012; 158: 824–34. doi: 10.1104/pp.111.185033 PMID: 22135431
12. Olukolu B a, Wang G-F, Vontimitta V, Venkata BP, Marla S, Ji J, et al. A genome-wide association study of the maize hypersensitive defense response identifies genes that cluster in related pathways. PLoS Genet. 2014; 10: e1004562. doi: 10.1371/journal.pgen.1004562 PMID: 25166276
13. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet. 2011; 43: 159–62. doi: 10.1038/ng.746 PMID: 21217756
14. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. Nature Publishing Group; 2013; 45: 43–50. doi: 10.1038/ng.2484

15. Nielsen R, Hubisz MJ, Clark AG. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 2004; 168: 2373–82. doi: 10.1534/genetics.104.031039 PMID: 15371362

16. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15: 1496–502. doi: 10.1101/gr.4107905 PMID: 16251459

17. Cheng AY, Teo Y, Ong RT. Assessing single nucleotide variant detection and genotype call- ing on whole-genome sequenced individuals. 2014; 1–7.

18. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. Nature Publishing Group; 2011; 12: 443–451. doi: 10.1038/nrg2986

19. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. BioMed Central Ltd; 2013; 5: 28. doi: 10.1186/gm432

20. Yu X, Sun S. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics. BMC Bioinformatics; 2013; 14: 274. doi: 10.1186/1471-2105-14-274

21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9: 357–9. doi: 10.1038/nmeth.1923 PMID: 22388286

22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; 00: 1–3.

23. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26: 589–95. doi: 10.1093/bioinformatics/btp698 PMID: 20080505

24. CLCbio website. Available: http://www.clcbio.com. Accessed 27 June 2013.

25. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013; 29: 2790–1. doi: 10.1093/bioinformatics/btt468 PMID: 23975764

26. SMALT. Available: http://www.sanger.ac.uk/resources/software/smalt/. Accessed 10 December 2013.

27. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011; 21: 936–9. doi: 10.1101/gr.111120.110 PMID: 20980556

28. Cheng AY, Teo YY, Ong RTH. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics. 2014; 30: 1707–1713. doi: 10.1093/bioinformatics/btu067 PMID: 24558117

29. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics. BioMed Central Ltd; 2012; 13 Suppl 8: S8. doi: 10.1186/1471-2164-13-S8-S8

30. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012; 1–9.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

32. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009; 25: 2283–2285. doi: 10.1093/bioinformatics/btp373 PMID: 19542151

33. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 2010; 26: i318–24. doi: 10.1093/bioinformatics/btq214 PMID: 20529923

34. Wang W, Wang P, Xu F, Luo R, Wong MP, Lam T-W, et al. FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data. Bioinformatics. 2014; 30: 2498–500. doi: 10.1093/bioinformatics/btu338 PMID: 24833803

35. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011; 39: e132. doi: 10.1093/nar/gkr599 PMID: 21813454

36. Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Boroevich K a, et al. A practical method to detect SNVs and indels from whole genome and exome sequencing data. Sci Rep. 2013; 3: 2161. doi: 10.1038/srep02161 PMID: 23831772

37. Park MH, Rhee H, Park JH, Woo HM, Choi BO, Kim BY, et al. Comprehensive analysis to improve the validation rate for single nucleotide variants detected by next-generation sequencing. PLoS One. 2014; 9. doi: 10.1371/journal.pone.0086664

38. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One. 2012; 7: e30087. doi: 10.1371/journal.pone.0030087 PMID: 22347999

39. Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011; 6: e19379. doi: 10.1371/journal.pone.0019379 PMID: 21573248

40. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, et al. RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun. Nature Publishing Group; 2013; 4: 2832. doi: 10.1038/ncomms3832

41. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. Nature Publishing Group; 2012; 44: 803–7. doi: 10.1038/ng.2313

42. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. Nature Publishing Group; 2014; 15: 335–46. doi: 10.1038/nrg3706

43. Botstein D, White RL, Skolnick M, Davis RW. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. 1980; 314–331.

44. Mascher M, Wu S, Amand PS, Stein N, Poland J. Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. 2013; 8: 1–11. doi: 10.1371/journal.pone.0076925

45. Homer N, Nelson SF. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol. BioMed Central Ltd; 2010; 11: R99. doi: 10.1186/gb-2010-11-10-r99

46. DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43: 491–8. doi: 10.1038/ng.806 PMID: 21478889

47. Lunter G. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. Bioinformatics. 2007; 23: i289–96. doi: 10.1093/bioinformatics/btm185 PMID: 17646308

48. Cartwright R a. Problems and solutions for estimating indel rates and length distributions. Mol Biol Evol. 2009; 26: 473–80. doi: 10.1093/molbev/msn275 PMID: 19042944

49. Bouchet S, Servin B, Bertin P, Madur D, Combes V, Dumas F, et al. Adaptation of maize to temperate climates: mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8) locus. PLoS One. 2013; 8: e71377. doi: 10.1371/journal.pone.0071377 PMID: 24023610

50. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, et al. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet. Public Library of Science; 2009; 5: e1000734. doi: 10.1371/journal.pgen.1000734

51. Stenson PD, Ball E V, Mort M, Phillips AD, Shiel J a, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003; 21: 577–81. doi: 10.1002/humu.10212 PMID: 12754702

52. Rincent R, Nicolas S, Bouchet S, Altmann T, Brunel D, Revilla P, et al. Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production. Theor Appl Genet. 2014; 127: 2313–31. doi: 10.1007/s00122-014-2379-7 PMID: 25301321

53. Rincent R, Moreau L, Monod H, Kuhn E, Melchinger AE, Malvar R a., et al. Recovering power in association mapping panels with variable levels of linkage disequilibrium. Genetics. 2014; 197: 375–387. doi: 10.1534/genetics.113.159731 PMID: 24532779

54. Zhang L, Li W-H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol. 2004; 21: 236–9. doi: 10.1093/molbev/msh010 PMID: 14595094

55. Akhunov E, Sehgal S, Liang H, Wang S, Akhunova A, Kaur G, et al. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol. 2012; doi: 10.1104/pp.112.205161

56. Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, et al. The physical and genetic framework of the maize B73 genome. PLoS Genet. 2009; 5: e1000715. doi: 10.1371/journal.pgen.1000715 PMID: 19936061

57. Clouse S.D., and Feldmann KA. Molecular genetics of brassinosteroid action. 1999; 163–190.

58. Almagro L, Gómez Ros L V, Belchi-Navarro S, Bru R, Ros Barceló a, Pedreño M a. Class III peroxidases in plant defence reactions. J Exp Bot. 2009; 60: 377–90. doi: 10.1093/jxb/ern277 PMID: 19073963

59. Hu J, Zhang Y, Wang J, Zhou Y. Glycerol affects root development through regulation of multiple pathways in Arabidopsis. PLoS One. 2014; 9: e86269. doi: 10.1371/journal.pone.0086269 PMID: 24465999

60. Morgulis A, Gertz EM, Schäffer A a, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006; 22: 134–41. doi: 10.1093/bioinformatics/bti774 PMID: 16287941

61. Ning Z, Cox AJ, Mullikin JC. SSAHA: A Fast Search Method for Large DNA Databases. 2001; 1725–1729. doi: 10.1101/gr.194201.1

62. Mace ES, Buhariwalla HK, Crouch JH. A High-Throughput DNA Extraction Protocol for Tropical Molecular Breeding Programs. 2003;

63. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, et al. Gramene 2013: comparative plant genomics resources. Nucleic Acids Res. 2013; 1–7. doi: 10.1093/nar/gkt1110 PMID: 24217918

64. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10: 421. doi: 10.1186/1471-2105-10-421 PMID: 20003500

65. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. 2010; 26: 841–842. doi: 10.1093/bioinformatics/btq033

66. Danecek P, Auton A, Abecasis G, Albers Ca, Banks E, DePristo Ma, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–8. doi: 10.1093/bioinformatics/btr330 PMID: 21653522

67. Haseneyer G, Schmutzer T, Seidel M, Zhou R, Mascher M, Schön C-C, et al. From RNA-seq to large-scale genotyping—genomics resources for rye (Secale cereale L.). BMC Plant Biol. 2011; 11: 131. doi: 10.1186/1471-2229-11-131 PMID: 21951788

68. Jia P, Li F, Xia J, Chen H, Ji H, Pao W, et al. Consensus rules in variant detection from next-generation sequencing data. PLoS One. 2012; 7. doi: 10.1371/journal.pone.0038470

69. Vergara I a, Frech C, Chen N. CooVar: co-occurring variant analyzer. BMC Res Notes. BMC Research Notes; 2012; 5: 615. doi: 10.1186/1756-0500-5-615 PMID: 23116482

70. Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. BMC Genomics; 2014; 15: 162. doi: 10.1186/1471-2164-15-162 PMID: 24571581

71. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19: 1639–45. doi: 10.1101/gr.092759.109 PMID: 19541911

72. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155: 945–59. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461096&tool = pmcentrez&rendertype = abstract PMID: 10835412

73. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 2007; 7: 574–578. doi: 10.1111/j.1471-8286.2007.01758.x PMID: 18784791

74. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164: 1567–87. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462648&tool = pmcentrez&rendertype = abstract PMID: 12930761

75. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14: 2611–20. doi: 10.1111/j.1365-294X.2005.02553.x PMID: 15969739

76. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008; 2008: 619832. doi: 10.1155/2008/619832 PMID: 18483572

77. Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, et al. Maize Metabolic Network Construction and Transcriptome Analysis. Plant Genome. 2013; 6: 1–12. doi: 10.3835/plantgenome2012.09.0025

Original Article

Cytogenetic and Genome Research

Cytogenet Genome Res 2014;142:66–78
DOI: 10.1159/000356460

Accepted: July 8, 2013
by M. Schmid
Published online: December 6, 2013

Kmasker – A Tool for in silico Prediction of Single-Copy FISH Probes for the Large-Genome Species *Hordeum vulgare*

T. Schmutzer L. Ma N. Pousarebani F. Bull N. Stein A. Houben U. Scholz

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Key Words
BAC-FISH · *Hordeum vulgare* · *k* -mer analysis · Physical mapping · Repetitive sequences · Single-copy FISH

Abstract
Specific localization of large genomic fragments by fluorescence in situ hybridization (FISH) is challenging in large-genome plant species due to the high content of repetitive sequences. We report the automated work flow (Kmasker) for in silico extraction of unique genomic sequences of large genomic fragments suitable for FISH in barley. This method can be widely used for the integration of genetic and cytogenetic maps in plants and other species with large and complex genomes if the probe sequence (e.g. BACs, sequence contigs) and a low coverage (8-fold) of unassembled sequences of the species of interest are available. Kmasker has been made publicly available as a web tool at http://webblast.ipk-gatersleben.de/kmasker.

Fluorescence in situ hybridization (FISH) is a very powerful tool for the integration of physical/genetic maps with the cytogenetic map of genomes. Contigs of overlapping bacterial artificial chromosome (BAC) clones could be anchored by FISH to chromosomes of plant species carrying relatively small genomes, such as *Oryza sativa* [Jiang et al., 1995], *Arabidopsis thaliana* [Fransz et al., 1998], *Sorghum bicolor* [Kim et al., 2002], *Brachypodium distachyon* [Hasterok et al., 2006], *Gossypium raimondii* [Wang et al., 2007], and *Solanum lycopersicum* [Szinay et al., 2008]. In large-genome species like barley, the direct mapping of entire BAC clones to chromosomes typically leads to a labeling of large parts or even the entire genome due to the presence of repetitive DNA elements that are highly conserved and dispersed throughout the genome.

Repetitive sequences of small-genome species are less frequent and often can be suppressed with Cot series fractions [Britten et al., 1974] allowing unique sequences to be detected by in situ hybridization. But, in large- and complex-genome species like barley, most BAC clones contain a low density of unique sequences and a high amount of different types of repetitive sequences [Steuernagel et al., 2009]. Therefore, cytological mapping of BACs in species with large genomes is more challenging. For instance, Zhang et al. [2004a, b] selected 56 RFLP-locus-specific BAC clones from the A and D genome donors of the hexaploid wheat for BAC-FISH. All labeled BACs resulted in FISH patterns similar to transposable

KARGER

E-Mail karger@karger.com
www.karger.com/cgr

1424–8581/13/1421–0066$38.00/0

Uwe Scholz
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)
Corrensstrasse 3
DE–06466 Gatersleben (Germany)
E-Mail scholz@ipk-gatersleben.de

elements or tandem repeats [Zhang et al., 2004a, b]. Suzuki and Mukai [2004] used 202 BACs of *Aegilops squarrosa* for FISH analysis on chromosomes of common wheat in the absence of unlabeled Cot-1 genomic DNA as competitor. Most of the BACs cross-hybridized with entire chromosomes as dispersed signals, and no BAC detected a single locus by FISH. So far, only a few barley BACs were cytologically mapped by preselecting clones encoding few repeats and using suppression hybridization [Lapitan et al., 1997; Stephens et al., 2004; Phillips et al., 2010]. Alternatively, due to the conservation and synteny of genes and diversity of repetitive sequences across species, it is possible to hybridize BAC clones from related species, such as *B. distachyon*. Ma et al. [2010] showed that only 2 of 13 *Brachypodium* labeled BACs gave signals in cross-species hybridization with barley chromosomes.

FISH technology has been improved in plants for the detection of small and single-copy fragments (few kilobases in length) [Kato et al., 2006; Danilova and Birchler, 2008]. Hence, an efficient way of identifying reliable low-copy sequences independent of the functional annotation of a sequence would expedite the routine development of FISH probes in physical mapping projects of plant genomes. Repeat masking is a crucial step in many sequence analysis applications like de novo assembly or annotation. If repetitive sequences from genomic sequences like BAC clones could be automatically identified with high reliability, it could be possible to develop a PCR-based strategy for the automated design of unique FISH probes. The identification of single-copy sequences is a computational challenge since each sub-sequence needs to be compared to the entire rest of the genome to estimate its occurrence. This has been hampered in large-genome species due to the absence of a reference sequence. Manually curated repeat databases like the Triticeae repeat composition TREP [Wicker et al., 2002] can be applied to mask known repeats based on sequence similarity. For this purpose, programs like RepeatMasker (http://www.repeatmasker.org) or DustMasker [Morgulis et al., 2006] can be applied. But, as Frith [2011] argued, these methods based on homology search can result in imperfect masking. Non-homologous similarity can confound the masking process and was observed mainly in low-complexity regions like extremely AT-rich sequences or CpG islands. But the main reason for incomplete masking is the incompleteness of repeat element databases.

The repetitiveness of DNA can also be determined mathematically by generating statistics of occurrence of a given sequence in a genome [e.g. Kurtz et al., 2008]. In *k*-mer analysis, the complexity of sequence comparison is reduced to a fixed length *k*. All *k*-mer sub-sequences occurring in the genome are analyzed, and their frequency is stored in an efficient data structure.

A high-quality reference sequence of the entire barley genome is not available, although a large amount of sequence information has become available recently by a number of next-generation sequencing approaches [Wicker et al., 2008, 2010; Mayer et al., 2009, 2011; International Barley Genome Sequencing Consortium, 2012]. *k*-mer counting does not require a draft or fully assembled genomic sequence. It is applied directly to raw reads, and for each observed *k*-mer, the *k*-mer index contains the sequence and frequency information, without the need of alignment to a reference sequence. *k*-mer approaches rely on exact sequence identity within a *k*-mer of length *k* and thus are sensitive to sequencing errors or slight differences between index and query sequence. However, with higher sequencing depth, the *k*-mer masking approach is less susceptible to sequencing errors inherent to all next-generation sequencing technologies. The repetitive nature of a genome compensates for the lack of sequencing depth, since most of the reads are non-unique and thus are present by multiple copies.

Here, we show that a *k*-mer index based on minimum 8-fold haploid genome sequencing depth enables routinely for the identification of low-copy sequences out of genomic sequence context suitable for the use as probes for single-copy FISH. We developed an automated work flow (Kmasker) for in silico extraction of unique genomic sequence information for the design of genomic single-copy FISH probes suitable to anchor BAC clones reliably onto barley chromosomes. This method can be widely used for the integration of genetic and cytogenetic maps in large and complex plant genomes.

Materials and Methods

k-mer Counting with Tallymer

Tallymer is a collection of programs for *k*-mer counting as well as for the construction of indices for large sequence datasets that was published by Kurtz et al. [2008]. Its efficiency is achieved by relying on enhanced suffix arrays to perform *k*-mer counting tasks instead of traditional approaches that applied hashing methods. *k*-mers are defined as substring of length *k* of a sequencing read over the nucleotide alphabet {A,C,G,T,N}, where N denotes the undetermined base. For all observed *k*-mers in the input dataset, the number of positions where a specific *k*-mer occurs is counted. In total, $n - k + 1$ *k*-mers are possible within each read of length *n* and *k*-mer size *k* (fig. 1). Within Tallymer, the routine 'mkindex' is used for counting and indexing *k*-mers based on a specified value of *k* in

Cytogenet Genome Res 2014;142:66–78
DOI: 10.1159/000356460

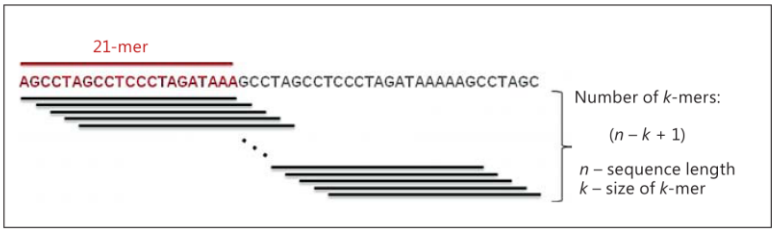

Fig. 1. *k*-mer. Visualization of *k*-mers that are defined as substrings of a given read sequence.

a dataset of sequence reads. The constructed index used in Kmasker is relying on 21-mers, where this size is chosen with respect to the uniqueness of *k*-mers. This uniqueness is taking into account those *k*-mers of the complete set which occur exactly once. Large values of *k* increase the number of these distinct *k*-mers but also increase the possibility to overlap an error that will sophisticate its frequency. Smaller *k* values have less specificity. We balanced both aspects and achieved the optimal information content for 21-mers. The 'suffixerator' routine of Tallymer is used to construct the enhanced suffix array that is required to search FASTA sequences.

In silico Prediction of Single-Copy FISH Probes

To detect single-copy sequences, we used the approach of mathematically defined repeats defined as *k*-mer frequency in a sequence dataset [Wicker et al., 2008]. In our analysis, we integrated into the *k*-mer index an 8-fold whole-genome shotgun (WGS) sequencing of the haploid barley genome. The 100-bp reads have been produced by Illumina GAIIx sequencing. The established *k*-mer index incorporated 42.8 Gb of sequence data (424 million reads) into an index structure of 323 gigabyte. This data is a subset of a recently published larger dataset (International Barley Genome Sequencing Consortium, 2012; EMBL ENA ERP001449, run ERR127105). In addition, we ran several experiments to study the impact of sequencing depth on prediction reliability. To obtain a set of genome indices with lower sequencing depth (6×, 4×, 2×, 1×, 0.5×, and 0.1×), we reduced the 8-fold WGS dataset by a random subsampling. Estimations of the information content in the constructed WGS data are shown in table 1.

Table 1. Sequence characteristics of the indexed read datasets

Barley WGS dataset	Index size, Gb	Sequence, Gb	Number of reads	Coverage[a]	Expected Lander-Waterman[b], %
0.1×	5	0.64	6,359,186	0.13×	11.8
0.5×	21	2.78	27,563,414	0.54×	42.1
1×	41	5.35	53,000,328	1.05×	65.9
2×	81	10.71	106,004,060	2.10×	87.7
4×	162	21.42	212,037,874	4.20×	98.5
6×	242	32.12	318,047,697	6.30×	99.8
8×	323	42.83	424,063,596	8.40×	99.98

We used the 8× dataset as a template and applied random selection to obtain the reduced datasets.

[a] The coverage estimation is based on a 5.1-Gb genome size of barley.

[b] The Lander-Waterman expectation estimates the percentage of the genome that is represented in the WGS sequencing [Lander and Waterman, 1988].

Introduction of the in silico Analysis Pipeline

Kmasker was implemented in Perl and is provided as web-accessible Common Gateway Interface (CGI) script. It integrates the Tallymer tool [Kurtz et al., 2008] for index construction and assessment of *k*-mer frequency. Post-processing procedures were implemented in Java Script using the d3-cloud (https://github.com/jasondavies/d3-cloud). PhantomJS is used for page automation to capture d3-cloud graphics (http://phantomjs.org/).

The pipeline has 3 main steps: (1) index construction, (2) index application and (3) sequence analysis. In the first step, reads that have been quality-trimmed by using pre-processing methods of CLC Assembly Cell v3.2.2 (http://www.clcbio.com/) are used to build an enhanced suffix array with a *k*-mer length of 21 bp. In the second step, sequences are virtually fragmented into overlapping 21-mers and each 21-mer is assessed for its frequency in the barley genome sequence index. The third step is evaluating the detected frequency values and masks repetitive positions based on user defined thresholds. A tentative functional annotation is computed by BLASTn [Altschul et al., 1997] comparison to public NCBI plant genomes (http://www.ncbi.nlm.nih.gov/) and to the repetitive sequence repository Repbase [Jurka et al., 2005]. Subsequently, the visualization module implemented in Kmasker is used to construct several graphical representation plots for all final results. The whole computational approach of the Kmasker pipeline is illustrated in figure 2. The in silico predicted sequences can be used as single-copy FISH probes for physical mapping in the large-genome species *Hordeum vulgare*.

Establishment of a Low-Complexity Control

We defined a control dataset to validate the detection of low-complexity and single-copy sequences. We used a published WGS assembly of barley cultivar Bowman [International Barley Genome Sequencing Consortium, 2012] and identified by read mapping regions of low copy number, assuming a read depth of 5 as maximum coverage. WGS contigs are available at EMBL/ENA under the accession numbers CAJX010000001–CAJX012077901. All mappings of paired end read data were performed with the Burrows-Wheeler Alignment tool (BWA, version 0.6.2) [Li and Durbin, 2010]. To further ensure uniqueness, we only used positions that are linked to exons [International Barley Genome Sequencing Consortium, 2012] and have adequate length for analysis (>200 bp). In total, this dataset has a cumulative length of 60.7 Mb.

Due to the stringency of the applied selection process for low complexity, we used the established dataset as a control to measure sensitivity and specificity of our prediction. To further estimate the validity of our prediction, we established a second control without prior selection for uniqueness including 91 published barley BAC clones covering 10.3 Mb of sequences [Steuernagel et al., 2009; Taudien et al., 2011]. Descriptive information on both control datasets is listed in table 2.

Sequence Dataset for in silico Evaluation

To demonstrate the general applicability and throughput of our methods on genome scale, we used a WGS dataset as well as genomic BAC assemblies published by the International Barley Genome Sequencing Consortium [2012]. The WGS contigs are available at EMBL/ENA (see above). The used BAC assemblies are available at NCBI GenBank under the accession numbers AC247243.1–AC247289.1, AC247294.1–AC250420.1, AC252611.1–AC253531.1, and AC250421.1–AC252610.1.

Primer Design and PCR Conditions

Primers were designed using the public software tool Batchprimer 3 [You et al., 2008]. Primer picking parameters were set as follows with the optimal Tm = 62°C (range 60–64°C with a difference of maximum 2°C between forward and reverse primers) and optimal GC content = 55% (range 45–60%). The minimum target size was 1,300 bp. Sequences that did not satisfy these conditions were rejected.

The PCR reagent mixture consisted of 1 µl of genomic DNA (5 ng/µl), 1 µl of 10× PCR buffer, 1 µl of dNTP mixture (2 mM each), 1 µl of primer mix (5 pmol/µl each), 0.05 µl of HotStar Taq DNA polymerase (Qiagen, Hilden, Germany), and 5.95 µl of ddH_2O. All fragments were amplified using the following touchdown PCR profile: an initial denaturing step of 15 min at 95°C was followed by 40 cycles with denaturation at 94°C for 30 s and extension at 72°C for 1 min. The annealing temperature was decreased in 1°C increments from 65°C in the first cycle to 60°C after the 5th cycle and was then kept constant for the remaining 35 cycles (always 30 s). After 40 cycles, a final extension step was performed at 72°C for 7 min. PCR amplifications were carried out using the GeneAmp PCR system 9700 (Applied Biosystems).

Preparation of FISH Probes

In silico-defined unique sequences were amplified by PCR using genomic DNA as template (primers are listed in table 3). Amplicons for the 5S ribosomal DNA (rDNA), which include the coding as well as the flanking spacer region, were generated by PCR as described by Fukui et al. [1994]. PCR products were purified using the QIAquick PCR Purification Kit (Qiagen) and directly labeled with Texas-red-dUTP (Invitrogen) or Alexa Fluor 488-5-dUTP (Invitrogen) by nick translation [Kato et al., 2006].

FISH

Chromosome preparations were obtained from ice water-treated fresh roots from barley cultivar 'Morex' as described earlier [Kato, 2011]. After treatment with 45% acetic acid and pepsin (Sigma-Aldrich; 0.1 mg/ml in 10 mM HCl) each for 10 min separately, chromosome preparations were post-fixed in 4% formaldehyde in 2× SSC for 10 min, dehydrated in an ethanol series (70, 90 and 96%) and air dried. The hybridization mixture contained 50% deionized formamide, 2× SSC, 1× TE, 50 ng/µl of single-copy probe, 10× ex-

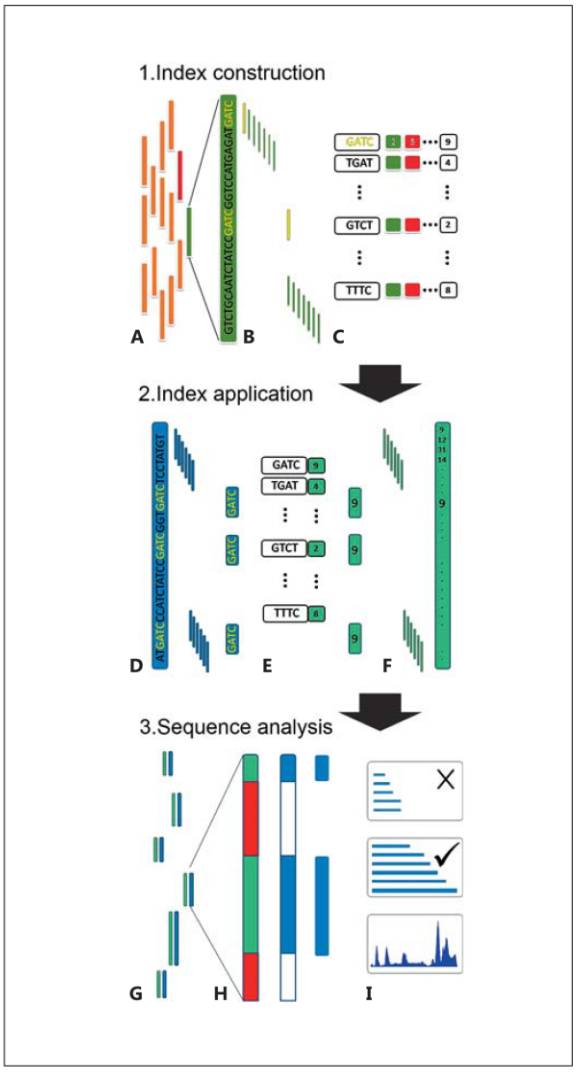

Fig. 2. Schema of the Kmasker pipeline. (1) Index construction for $k = 4$. Each WGS read (**A**) is split in overlapping k-mers (**B**). A frequency table for all occurring k-mers is calculated by repeating this step for all reads (**C**). (2) Index application. For each overlapping k-mer of the query sequence shown in blue (**D**), the frequency is provided by the index (**E**). All values are combined in a frequency representation for the query sequence shown in green (**F**). (3) Sequence analysis. The index is applied to all query sequences to calculate their related frequency representation (**G**). The representation is evaluated to search for low-copy sequences shown in green and masked high-frequency regions shown in red (**H**). By applying a size threshold to all low-copy regions, a final set of candidate sequences is selected for FISH-derived primer design (in blue). Sequences with low repetitive complexity and length >1 kb are targets for a successful primer design applicable for FISH (**I**).

Table 2. Sequence characteristics of the analyzed query datasets

Dataset	Number of contigs	Contig size			Nucleotide content, %			
		total	average, bp	L50, bp	A	C	G	T
FISH[a]	38	257.8 kb	6,784	10,325	0.282	0.219	0.215	0.283
Control (BAC)[b, c]	1,511	10.3 Mb	6,843,455	41,767	0.278	0.223	0.224	0.275
Control (exon)[c]	101,509	60.7 Mb	598	770	0.251	0.247	0.250	0.252

The table lists general statistics for the 3 datasets to which constructed *k*-mer indices have been applied.
[a] From the FISH dataset, we selected single-copy sequences that were integrated into our FISH analyses.
[b] The dataset of 91 BACs fragmented in 1,511 contigs was used as independent control to study the impact of sequencing depth in a large dataset.
[c] The BAC and exon datasets were used as control to validate low-copy predictions.

Table 3. Summary of sequences used for probe development

Fingerprinted contig ID[a]	Genetic position, cM	Morex_assembly3_contig[a]	BAC to which the Morex contig is assigned	Fingerprinted contig size, kb	Running number	Amplified size on genomic DNA, bp	Total size of FISH probes, bp	Primer sequences (forward/reverse)
678	140.7932	morex_contig_37676	HVVMRXALLE-A0297L21	364.56	7_1	1,496	4,399	TGAGGTACAATACCTGCTCAACG/ GAATAGCGTGGCTCCAATCATAG
678	140.7932	morex_contig_2553533	HVVMRXAL-LEA0258M17	364.56	7_2	1,332		CTCTCATCGGTGCTCAGTGG/ CCCAGGTTCCTTCTCAACCAT
678	140.7932	morex_contig_2553533	HVVMRXAL-LEA0258M17	364.56	7_3	1,571		AGCTTAGCTGACTTAGGGCCAGT/ GCATACGCTGAGAGAAGTTACCC
38863	135.6232	morex_contig_5743	HVVMRXALLE-A0330N02	1,125.92	5_1	2,346	5,065	GAATGTTCCTGCTGTTTGTGCT/ AATCGTTCTAATGCGCACAGC
38863	135.6232	morex_contig_5743	HVVMRXALLE-A0330N02	1,125.92	5_2	2,719		AACTTCCAAAGGTCGTTTCCAC/ CCCATCTCGTTTCGATCTGTTT

[a] According to the International Barley Genome Sequencing Consortium [2012].

cess of sonicated salmon sperm DNA, and 2 ng/μl 5S rDNA probe. The hybridization mixture and the treated chromosome slides were denatured together on a heating plate at 80°C for 2 min and incubated in a moist chamber at 37°C overnight. Post-hybridization washing was done in 2× SSC for 20 min at 55°C. After dehydration in an ethanol series (70, 90 and 96%), the slides were air dried at room temperature and counterstained with 4′,6-diamidino-2-phenylindole (DAPI) in Vectashield (Vector Laboratories). Images were taken using a cooled CCD camera (Spot 2e, Diagnostic Instruments) and an epifluorescence microscope (Axioplan 2, Zeiss) with a Plan Apochromat 63×/1.40. The images were pseudocolored and merged using Adobe Photoshop. The chromosome nomenclature of the barley genome is according to the internationally agreed recommendation [Linde-Laursen et al., 1997].

Results and Discussion

The correct identification of repeat-free sequences is instrumental to the generation of FISH probes suitable for unique cytogenetic mapping of BAC contigs or other large-insert DNA fragments derived from species with large genomes. We addressed this challenge by the application of the *k*-mer analysis which uses unassembled sequence reads obtained by Roche/454 pyrosequencing, Illumina (Solexa) sequencing, or any other high-throughput sequencing technology.

Evaluation of Single-Copy Predictions

To evaluate accuracy of a Kmasker prediction for low-copy regions, we applied our pipeline to the established control datasets and cross-checked their prediction with a coverage estimation obtained by BWA mapping. For the constructed exon, control dataset uniqueness (TP = true positive) was assigned when less than 2 reads were mapped to a position; otherwise we considered the position as non-unique (TN = true negative). By this we established a dataset to estimate the sensitivity (TP/(TP + FN)) and specificity (TN/(TN + FP)), where FN = false negative and FP = false positive. We could verify the

Table 4. Prediction errors

k-mer frequency[a]	Prediction error[b], %	*k*-mer frequency[a]	Prediction error[b], %	*k*-mer frequency[a]	Prediction error[b], %
0	40.23	30	5.67	200	4.69
1	14.36	50	5.34	300	4.44
2	11.37	60	5.23	400	4.32
5	9.73	70	5.14	500	4.20
10	6.87	80	5.07	700	4.12
15	6.66	90	5.01	900	4.02
20	6.03	100	4.97	1,000	3.95

Comparison of a *k*-mer analysis of a 0.1× index to an 8× reference index. 10.3 Mb of BAC sequences have been analyzed.

[a] The '*k*-mer frequency' column indicates which *k*-mer values are considered for comparison.

[b] The column 'Prediction error' is giving the percentage of bases for its corresponding *k*-mer frequency that have significant difference between the prediction of the 0.1× and the reference index (8×).

uniqueness predicted by the *k*-mer approach for 97.16% of positions (TP) where only 2.84% of unique positions missed a correct classification (FN), assigning a sensitivity of 98.44% to our methods. Furthermore, a specificity of 99.78% was reached when considering positions classified as non-unique by the mapping approach (TN), but predicted as unique by the *k*-mer approach (FP). The low number of erroneously called unique positions (1.49%) was further reduced to 0.01% when applying a parameter that requires a minimal length when classifying regions as unique. Assuming a putative sequence error, a *k*-mer frequency will decrease at these positions and might cause a false prediction of uniqueness. The length of sequencing errors is expected to be a few base pairs only. Referring to the study of Balzer et al. [2011] who systematically explored sequencing errors, we applied a 50-bp threshold to cover those erroneous positions and observed a significant increase of reliability. In addition, we evaluated the second control set consisting of 91 barley BAC sequences to estimate the performance in an unbiased dataset. *k*-mer methods predicted 2.95% of the analyzed 10.3 Mb to be unique where 95.25% of them can be verified by the mapping approach. The presented *k*-mer methods are able to detect unique and low-copy regions with high sensitivity and specificity. Therefore, we applied the constructed index (table 1) with a stringent setting (100-bp size threshold, *k*-mer frequency <3) to genomic barley sequences (table 2) to detect single-copy sequences that subsequently were integrated into FISH analyses.

To study the impact of sequencing depth for the correct masking of repetitive sequences and reliable prediction of single-copy sequences, we tested different levels of sequencing depth in WGS datasets. For the evaluation, we used a dataset of 91 barley BACs [Steuernagel et al., 2009; Taudien et al., 2011] and independently applied all constructed indices with identical parameter settings.

Previous studies that also aimed at the detection of repetitive DNA on the basis of sequence statistics [Wicker et al., 2008; Ma et al., 2010] referred as well to the concept of *k*-mer analysis but with extremely low-coverage sequencing. Here, we want to emphasize the importance of sequencing depth for FISH probe development. We applied the extremely low-coverage approach using a 0.1-fold coverage WGS dataset and directly compared the estimations to the 8-fold WGS dataset to evaluate reliability and error rates in the prediction. Therefore, we analyzed low-coverage (<5) positions in the 10.3-Mb BAC dataset. In 9.73%, we observed wrong interpretation by applying the 0.1-fold index. In contrast, using an index of 4-fold WGS coverage, even in low-coverage elements the prediction is correct in 98.7% of the targets. Table 4 presents the complete results of that analysis. Positions that are highly repetitive can be reliably predicted even by applying a low-coverage WGS index, whereas rare or single-copy elements need an index of at least 8-fold WGS coverage for robust predictions.

Improvement of the Prediction Accuracy

In addition, we observed positions that are not covered by the index, which we further refer to as positions omitted by index (POI). When comparing corresponding sequence positions that have been analyzed with indices of different read depth, we hypothesize that an increase of sequencing depth positively affects the results by decreasing the number of POI.

To assess the importance of sequencing coverage, we also evaluated the information gain that is obtained when increasing the number of sequences. We evaluated this criterion by monitoring the amount of POI. Therefore, we compared results of consecutive indices (0.1× vs. 0.5×; 0.5× vs. 1×...). The proportion of POI in 10.3 Mb of sequences dropped from 49.24% (0.1×) to 9.78% (8×) as shown in table 5. The substantial gain of information is achieved up to the 4-fold barley index. It already corrected 70% of the POI observed in the 0.1-fold index. Further 10% can be corrected when applying an 8-fold index. It was expected that the increase of sequence information positively affects reliability of the prediction [Schatz et al., 2012], and with our study we show that even on a minimal depth of 8-fold sufficient sensitivity is observed.

The origin of *k*-mer sequences omitted by an index can be due to various reasons. Unique sequences have a low-

Table 5. Index statistics

Dataset	POI	% POI	Information gain	Candidates FISH (>3 kb)	Sensitivity	Average *k*-mer frequency	P50 median[a]	P75[a]
0.1×	5,091,223	49.24	49.24	3,391	20.4	227	1	56
0.5×	3,891,139	37.63	11.61	2,470	28.0	985	2	245
1.0×	3,185,511	30.81	6.82	676	91.0	1,899	4	470
2.0×	2,333,090	22.56	8.24	692	93.4	3,796	8	940
4.0×	1,530,101	14.80	7.77	699	94.3	7,587	16	1,882
6.0×	1,181,023	11.42	3.38	719	96.5	11,381	25	2,821
8.0×	1,011,104	9.78	1.64	716	96.6	15,177	33	3,762

Evaluation of information content for different constructed indices that were applied to 91 BACs. To investigate the gain of information that can be achieved by increasing the sequencing depth, we calculated sequence positions where the index could not assign a frequency (POI). To estimate the completeness, we considered the masked 8× dataset as a reference and analyzed the increase of information gain for each index.

[a] P50 and P75 values are percentiles for the respective indexes.

er possibility to be sequenced and thus are not represented in the index, especially when considering sequencing errors or ploidy [Schatz et al., 2012]. This lack of information consequently leads to a higher possibility that a *k*-mer of the query sequence (e.g. BAC) is not captured in the index. This can be expressed by the Lander-Waterman expectation [Lander and Waterman, 1988] shown in table 1. Thus, when evaluating the impact of sequencing depth on predictability, the substantial drop of positions without an assigned frequency is correlated to an increasing Lander-Waterman score.

As previously shown by Novak et al. [2010] and Wicker et al. [2010], low-pass sequencing (like 0.1× WGS) is sufficient to predict highly repetitive regions of a genome by a graph-based cluster approach or *k*-mer analysis. With the presented study, we extend our perspective to low-copy sequences. As we showed with our observations regarding POI, the precision to estimate low *k*-mer frequencies is imperfect when applying indices originating from low-coverage sequencing. In consequence, the required completeness of *k*-mer masking is not sufficient, especially for repetitive elements that occur in lower frequency within a genome.

To illustrate that low-copy sequences represent a substantial amount of *k*-mer in the constructed index, we investigated the percentile of 21-mers for the barley genome (fig. 3). We assessed those *k*-mers that align to the 10.3-Mb BAC dataset and observed for ~30% a *k*-mer frequency <30 (~50% <400). This emphasizes the need of adequate sequencing coverage, required to capture those regions with sufficient reliability.

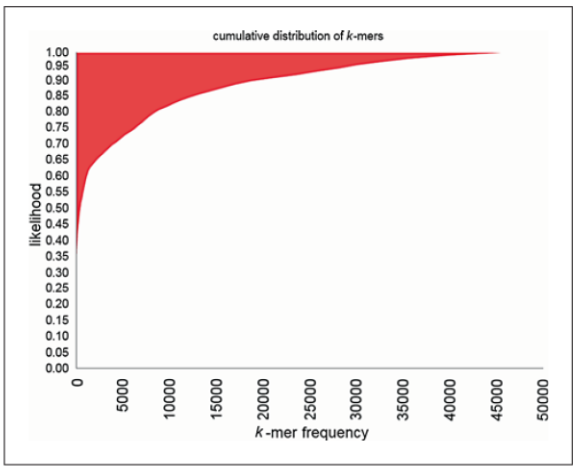

Fig. 3. Cumulative distribution of 21-mers in the barley genome. For all 21-mer sequences in the analyzed datasets, we captured their proportion in the barley genome and constructed a cumulative distribution of 21-mers in the barley genome. The graphic presents the possibility for each *k*-mer frequency.

The frequency plot shown in figure 4 further illustrates the impact of sequencing depth. It visualizes the results we achieved when applying different indices to an exemplary BAC sequence. According to our expectations, the precision of the *k*-mer approach is sufficient in frequency peaks, but in rare or unique sequences (indicated by blue bar in fig. 4) the permissive character of a low sequencing index is not strict enough. Here, higher sequencing is re-

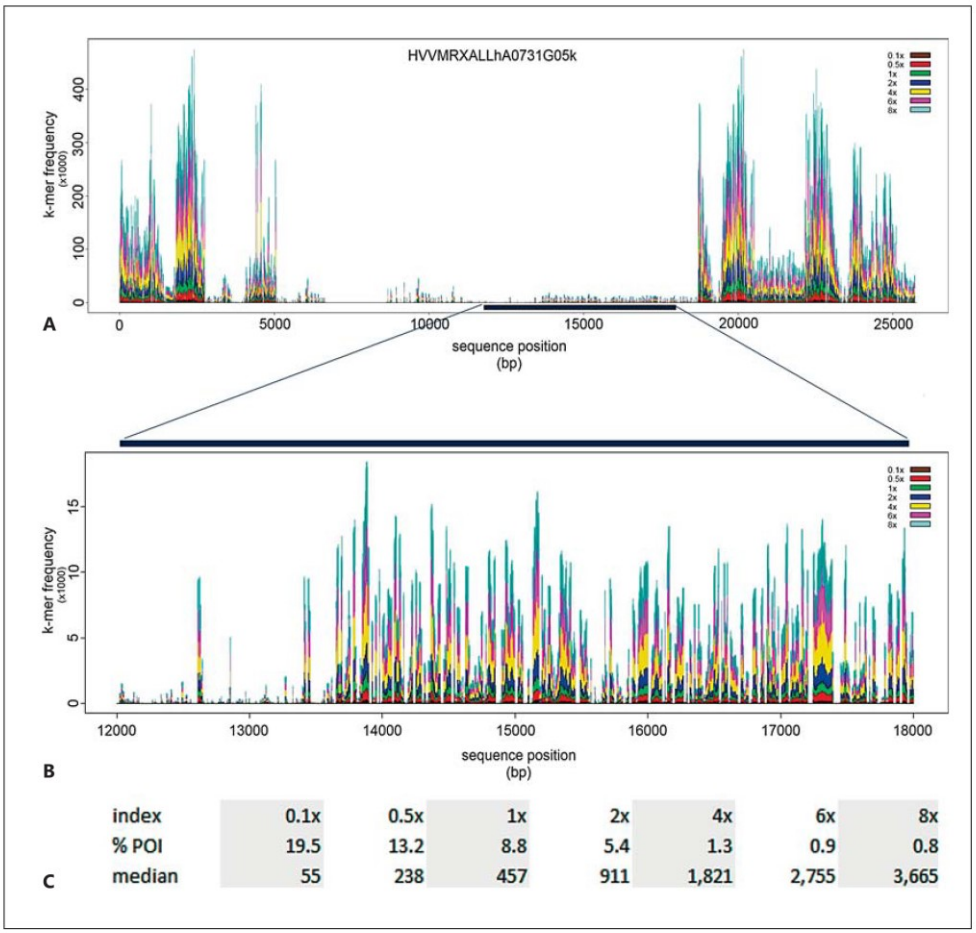

Fig. 4. Impact of sequencing depth on *k*-mer prediction. **A** The graphic presents all 21-mer counts for each sequence position of a barley BAC comprising a length of 25.8 kb (HVVMRXALLhA0731G05k). Colors refer to the 7 different indices that have been applied independently. For each sequence position (x-axis), their height is reflecting the *k*-mer frequency within the barley genome (y-axis). **B** Detailed view into lower frequency perspective of the BAC (blue bar at 12–18 kb). **C** Statistical details. For different indices, rates of position without frequency assignment (% POI) are presented including the median *k*-mer frequency observed in this BAC.

quired to reach sufficient reliability. When applying an 8-fold index, 99.2% of the BAC sequence is covered by *k*-mers, whereas 19.5% are not covered when applying an 0.1-fold index. This indicates that to enrich the detection of unique sequences should not consolidate a low-coverage sequencing index, because this might lead to erroneous prediction.

To investigate the sensitivity of repeat masking by the applied *k*-mer index approach, we compared the results obtained by applying each of the produced indices. As a reference, we used the masking results obtained by the 8-fold WGS coverage index plus an additional homology check against a database with annotated Triticeae repeat sequences [Wicker et al., 2002] (http://wheat.pw.usda.gov/ITMI/Repeats, TREP). 94.3% of frequently occurring repetitive elements can be detected in silico by applying an index that has been constructed based on a 4-fold WGS coverage read dataset (97.7% when applying TREP

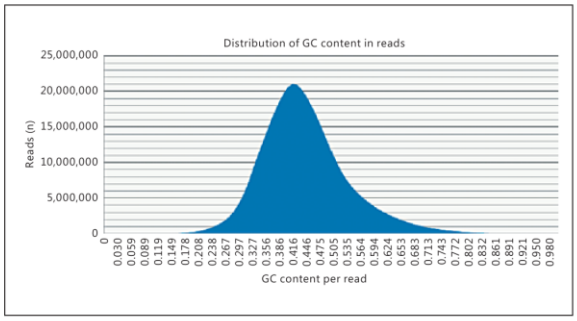

Fig. 5. Distribution of GC content. The graphic shows the distribution of GC content per read observed in the read set which was integrated into the *k*-mer index.

control subsequently to the *k*-mer masking). Nevertheless, like in genome projects where sequencing with low coverage is not recommended [Alkan et al., 2011], we want to emphasize too that further increase of sequencing depth can reduce the rate of false positive single-copy sequences. For detection of single-copy sequences, we observed sufficient masking by applying an 8-fold WGS dataset. Nevertheless, the combination with a homology search to repeat databases should be considered to mask omitted positions. Furthermore, deeper sequencing still can increase reliability of a prediction, especially in low-copy regions. But the masking quantity is less affected by the additional gain of information. The increase of sequence information positively affects reliability, but on the other side is requiring more storage and computational power. Following the progressive requirement for storage which we presented in table 1, an index constructed from a 50-fold WGS of barley would require ~2 TB of storage.

Various sequencing biases are described in the literature in terms of Illumina sequencing, known as strand bias [Guo et al., 2012] or GC content bias [Benjamini and Speed, 2012]. While strand biases are more critical in diversity aspects and might tend to call false positive SNPs, the GC content bias can affect a *k*-mer-based prediction of low-copy sequences. Major contribution to this effect is the PCR amplification step that can lead to an underrepresentation or absence of loci with extreme base composition. Different protocols have been published to correct this GC content bias [Aird et al., 2011; Benjamini and Speed 2012]. We analyzed the WGS dataset that has been integrated into the *k*-mer index for its GC content per read and found no clear indication that GC-rich genome segments are absent (fig. 5). Thus, we expect only a minor influence of sequencing bias on our methods. Nevertheless, to minimize a potential impact of the GC content bias, we encourage the user to double-check a repeat masking with homology-based methods using repositories of repeat libraries (TREP, Wicker et al. [2002]; Repbase, Jurka et al. [2005]).

Finally, we show the general applicability and throughput of our methods on a genome scale. Therefore, we ran Kmasker on 2 large genomic datasets (WGS contigs and BAC assemblies) published by the International Barley Genome Sequencing Consortium [2012]. These genomic datasets contain sequence contigs of 6,278 barley BACs (789.8 Mb) and a complete WGS assembly of barley cultivar Bowman (1.8 Gb). For 45% of the 6,278 BACs, we could identify low-copy regions of required size (>3 kb) that together comprised 18.1 Mb. Analysis of the assembled WGS contigs provided additional 44.7 Mb of sequences (10,976 WGS contigs) usable for FISH probe design. These automatically detected target regions are directly suitable for FISH and therefore can be widely used for the integration of genetic and cytogenetic maps.

Verification of in silico Predicted Single-Copy FISH Probes

To demonstrate the functionality of in silico predefined single-copy probes, barley genomic sequence information associated with 2 selected fingerprinted contigs (BAC contigs 678 and 38863, 0.36 and 1.13 Mb in size, respectively) was used for FISH probe development and was processed by Kmasker (exemplary data shown in fig. 6). For each contig, unique sequences were identified and corresponding PCR products (ranging from 1.3 to 2.7 kb) were generated and pooled to produce FISH probes with a cumulative length between 4.4 and 5.0 kb (table 3). A 5S rDNA-specific probe was used to distinguish the 7 chromosome pairs of barley [Fukui et al., 1994]. Eight out of 9 contig-specific single-copy probes yielded specific signals on chromosome 2H even without using Cot-DNA to block high-copy repeats (fig. 7). In a few cases, a tolerable threshold level of background signals was present, but the specific signals were unambiguously visible at the expected chromosome 2HL. Contig 678 was genetically anchored to the distal end of chromosome 2HL (fig. 7A). Contig 38863, which was allocated to a different region, hybridized to the distal end of chromosome 2HL as well (fig. 7B). Kmasker removed most of the repetitive sequences in the analyzed WGS contig and BAC sequences, thus the obtained signals present the true physical location of the corresponding BACs or WGS

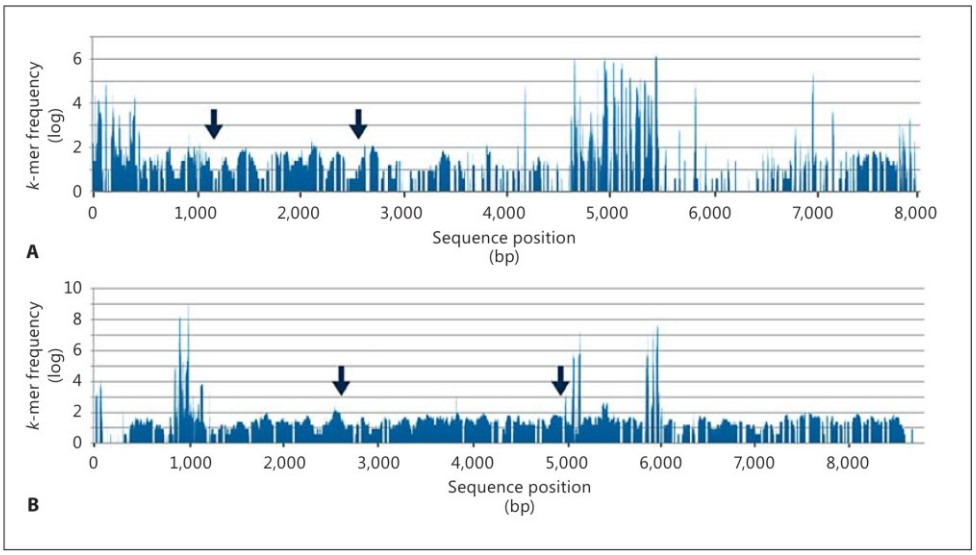

Fig. 6. *k*-mer frequency and primer positions of fingerprinted contigs 678 (**A**) and 38863 (**B**). Arrows indicate positions of primers used in the FISH experiment. Frequency values have been logarithmically transformed.

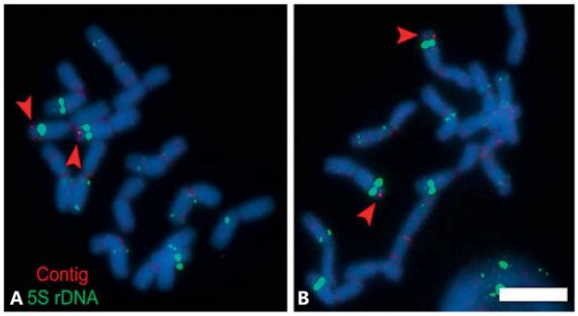

Fig. 7. FISH with selected unique sequence pools from the fingerprinted contigs. Pooled probe from contig 678 (**A**, in red) and contig 38863 (**B**, in red) can be visualized on somatic chromosome spreadings. Red arrowheads point to the signals. 5S rDNA is labeled in green. Bar = 10 µm.

contigs, in contrast to the previously dispersed signals on entire chromosomes obtained when performing FISH analysis with BAC clones in barley [Suzuki and Mukai, 2004]. The only requirement for the successful detection of the FISH probe is to generate probe sizes longer than the lower limit of FISH detection sensitivity (roughly 3 kb) using current technology [Kato et al., 2006; Danilova and Birchler, 2008; Ma et al., 2010]. One of our probes failed to produce a clear signal and was not considered for further optimization due to lack of enough fluorescent signal intensity.

Kmasker – A Web-Based Tool for Masking of Repeated Sequences

Our methods for in silico detection of single-copy sequences are based on the tool Tallymer for efficient index construction and access to these index structures [Kurtz et al., 2008]. These methods have been integrated into our publicly available web-based pipeline (http://webblast.ipk-gatersleben.de/kmasker) to identify high-quality candidates that are unique or occur at very low frequency in the genome. To compute *k*-mer counts, Kmasker is using an index that consists of ~424 million single-end reads. These were obtained as a subset from the published barley dataset (International Barley Genome Sequencing Consortium [2012]; EMBL ENA ERP001449, run ERR127105). Using the web-interface, users can analyze input sequences of a maximum length of 150 kb. Details on parameters, usage and output files are explained in the online help section that guides the user to run the tool in 5 steps. A complete example of in-

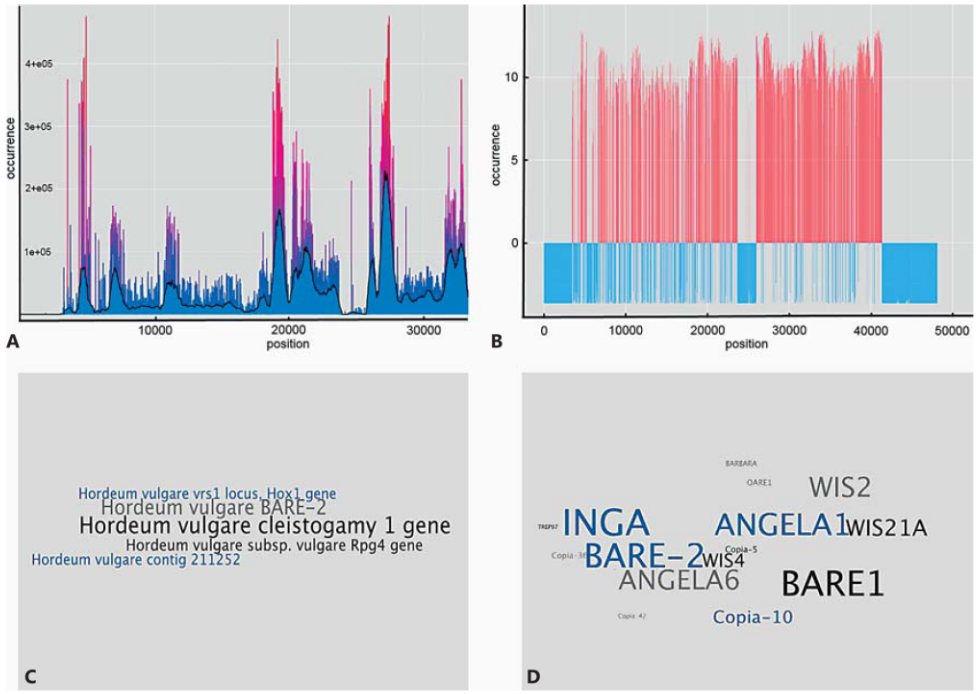

Fig. 8. Kmasker visualization. Graphical representation for barley BAC HVVMRXALLhA0731G05. **A** Frequency plot. For each sequence position, a *k*-mer frequency is calculated. **B** Graphic to determine segments of over- (red) and underrepresented (blue) *k*-mers. For visualization aspects, a logarithmic scaling is applied (log10 and natural log, respectively). **C** Annotation cloud. Graphical visualization of similarity to NCBI nucleotide collection detected by BLASTn. **D** Repeat cloud. Graphical visualization of similarity to repeats listed in Repbase (gras) detected by BLASTn.

put and output data, a version history, fields of application, and a section with frequently asked questions (FAQ) can be found at the webpage to support the community when using Kmasker. The masking process is influenced mainly by the repeat threshold which the user can adjust to define the tolerated repetitiveness. The second parameter ('size threshold') sets the size of contigs that will be included in the results file after removing the repetitive elements. All contigs below the provided threshold are discarded.

As a post-processing step, we incorporated a module into Kmasker for automated visualization of final results. In the current version, 4 main graphics are generated. The first panel (fig. 8A) illustrates the *k*-mer frequency for each position of the input sequence. The second graphical output (fig. 8B) illustrates whether a *k*-mer is over- or underrepresented when comparing to the theoretical expectation of a 21-mer in a genome of 5.1 Gb. With the third and fourth graphical output, we provide a tentative functional annotation visualized by using the method of tag clouds [Schrammel et al., 2009]. Similarity is assessed by comparing to the public sequence database (http://www.ncbi.nlm.nih.gov/) using BLAST analysis [Zhang et al., 2000]. We therefore extracted all sequences of the Triticeae tribe and performed a BLASTn comparison (E-value <1.0E-05). In addition, we evaluated segments of the user input sequence that are classified as repetitive by Kmasker depending on user-defined thresholds. By comparing to the reference database of eukaryotic repetitive DNA [Jurka et al., 2005], we provide a conservative annotation of detected repetitive elements. To transform these annotations into a tag cloud, we evaluated the description of all best hits and used their corresponding hit length to calculate a weighted list. These weights are used to directly transform significant similarities into visual tags where large fonts represent significant terms and small fonts less

significant terms. This allows the user to easily investigate sequence information (fig. 8C, D).

Users will receive all results via email including graphical visualization. In addition to designing single-copy FISH probes, our masking approach is applicable also for analyses in which repetitive sequences inhibit efficient primer design or sequence diversity study of the barley genome.

A major limitation of *k*-mer analysis is the dependency on complete sequence identity. Therefore, Kmasker is not suitable in cross-species applications, and constructed indices should be applied only to the species from which they originate. Even within groups of closely related species (e.g. grasses), the level of conservation on a whole-genome scale is too low to apply indices among different species. Thus, only highly conserved homologous sequences can be identified. This negatively impacts the performance of our methods, and we recommend to use only barley sequences within Kmasker to ensure valid masking. Our aim is to integrate WGS data of other plant genomes. Furthermore, we will actively maintain the application and provide further tools for downstream analysis.

In summary, we present the tool Kmasker which was designed for the prediction of highly efficient, single-copy FISH probes for cytogenetic (physical) mapping in the genome of barley and demonstrate the power of our methods by practical examples. With our study, we present the first web-based tool for in silico repeat masking in barley. Under various experimental conditions, we examined the reliability and emphasized the importance of sequencing depth in *k*-mer predictions. Results of this work are integrated into a pipeline to detect single-copy sequences and thus will accelerate and increase the success rate of FISH experiments in barley. However, the same approach can easily be adapted to any other large-genome species whenever sufficient whole-genome sequence information and larger contiguous target sequences are available for physical mapping.

Acknowledgements

We would like to thank Thomas Münch for his assistance in developing the Kmasker web version. We also thank Natasha Glover, Sebastian Beier and Martin Mascher for proofreading of the manuscript. This project was supported by the Leibniz Association (WGL) in the context of the 'Pakt für Forschung und Innovation/WGL' and the German Federal Ministry of Education and Research (BMBF) in the frame of the project BARLEX (FKZ 0314000A).

References

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al: Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12:R18 (2011).

Alkan C, Sajjadian S, Eichler EE: Limitations of next-generation genome sequence assembly. Nat Methods 8:61–65 (2011).

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402 (1997).

Balzer S, Malde K, Jonassen I: Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 27:i304-i309 (2011).

Benjamini Y, Speed TP: Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40:e72 (2012).

Britten RJ, Graham DE, Neufeld BR: Analysis of repeating DNA sequences by reassociation. Methods Enzymol 29:363–405 (1974).

Danilova TV, Birchler JA: Integrated cytogenetic map of mitotic metaphase chromosome 9 of maize: resolution, sensitivity, and banding paint development. Chromosoma 117:345–356 (2008).

Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G: Cytogenetics for the model system *Arabidopsis thaliana*. Plant J 13:867–876 (1998).

Frith MC: A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res 39:e23 (2011).

Fukui K, Kamisugi Y, Sakai F: Physical mapping of 5S rDNA loci by direct-cloned biotinylated probes in barley chromosomes. Genome 37:105–111 (1994).

Guo Y, Li J, Li CI, Long J, Samuels DC, Shyr Y: The effect of strand bias in Illumina short-read sequencing data. BMC Genomics 13:666 (2012).

Hasterok R, Marasek A, Donnison IS, Armstead I, Thomas A, et al: Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. Genetics 173:349–362 (2006).

International Barley Genome Sequencing Consortium: A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716 (2012).

Jiang J, Gill BS, Wang GL, Ronald PC, Ward DC: Metaphase and interphase fluorescence in situ hybridization mapping of the rice genome with bacterial artificial chromosomes. Proc Natl Acad Sci USA 92:4487–4491 (1995).

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467 (2005).

Kato A: High-density fluorescence in situ hybridization signal detection on barley (*Hordeum vulgare* L.) chromosomes with improved probe screening and reprobing procedures. Genome 54:151–159 (2011).

Kato A, Albert PS, Vega JM, Birchler JA: Sensitive fluorescence in situ hybridization signal detection in maize using directly labeled probes produced by high concentration DNA polymerase nick translation. Biotech Histochem 81:71–78 (2006).

Kim JS, Childs KL, Islam-Faridi MN, Menz MA, Klein RR, et al: Integrated karyotyping of sorghum by in situ hybridization of landed BACs. Genome 45:402–412 (2002).

Kurtz S, Narechania A, Stein JC, Ware D: A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics 9:517 (2008).

Lander ES, Waterman S: Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239 (1988).

- Lapitan NLV, Brown SE, Kennard W, Stephens JL, Knudson DL: FISH physical mapping with barley BAC clones. Plant J 11:149–156 (1997).
- Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595 (2010).
- Linde-Laursen I, Heslop-Harrison JS, Shepherd KW, Taketa S: The barley genome and its relationship with the wheat genomes. A survey with an internationally agreed recommendation for barley chromosome nomenclature. Hereditas 126:1–16 (1997).
- Ma L, Vu GT, Schubert V, Watanabe K, Stein N, Houben A, Schubert I: Synteny between *Brachypodium distachyon* and *Hordeum vulgare* as revealed by FISH. Chromosome Res 18:841–850 (2010).
- Mayer KFX, Taudien S, Martis M, Simková H, Suchánková P, et al: Gene content and virtual gene order of barley chromosome 1H. Plant Physiol 151:496–505 (2009).
- Mayer KFX, Martis M, Hedley PE, Hana S, Steuernagel B, et al: Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263 (2011).
- Morgulis A, Gertz EM, Schäffer AA, Agarwala R: A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13:1028–1040 (2006).
- Novák P, Neumann P, Macas J: Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378 (2010).
- Phillips D, Nibau C, Ramsay L, Waugh R, Jenkins G: Development of a molecular cytogenetic recombination assay for barley. Cytogenet Genome Res 129:154–161 (2010).
- Schatz MC, Witkowski J, McCombie WR: Current challenges in de novo plant genome sequencing and assembly. Genome Biol 13:243 (2012).
- Schrammel J, Deutsch S, Tscheligi M: Visual search strategies of tag clouds – results from an eyetracking study, in Gross T, Gulliksen J, Kotzé P, Oestreicher L, Palanque P, et al (eds): Human-Computer Interaction – INTERACT 2009, pp 819–831 (Springer, Berlin 2009).
- Stephens JL, Brown SE, Lapitan NL, Knudson DL: Physical mapping of barley genes using an ultrasensitive fluorescence in situ hybridization technique. Genome 47:179–189 (2004).
- Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, et al: De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. BMC Genomics 10:547 (2009).
- Suzuki G, Mukai Y: Plant BAC libraries as tools for molecular cytogenetics, in Williams CR (ed): Focus on Genome Research, pp 195–210 (Nova Science Publishers, New York 2004).
- Szinay D, Chang SB, Khrustaleva L, Peters S, Schijlen E, et al: High-resolution chromosome mapping of BACs using multi-colour FISH and pooled-BAC FISH as a backbone for sequencing tomato chromosome 6. Plant J 56: 627–637 (2008).
- Taudien S, Steuernagel B, Ariyadasa R, Schulte D, Schmutzer T, et al: Sequencing of BAC pools by different next generation sequencing platforms and strategies. BMC Res Notes 4:411 (2011).
- Wang K, Guo W, Zhang T: Detection and mapping of homologous and homoeologous segments in homoeologous groups of allotetraploid cotton by BAC-FISH. BMC Genomics 8:178 (2007).
- Wicker T, Matthews DE, Keller B: TREP: a database for Triticeae repetitive elements. Trends Plant Sci 7:561–562 (2002).
- Wicker T, Narechania A, Sabot F, Stein J, Vu GT, et al: Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics 9:518 (2008).
- Wicker T, Buchmann JP, Keller B: Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res 20: 1229–1237 (2010).
- You FM, Huo N, Gu YQ, Luo MC, Ma Y, et al: BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:253 (2008).
- Zhang P, Li WL, Friebe B, Gill BS: Simultaneous painting of three genomes in hexaploid wheat by BAC-FISH. Genome 47:979–987 (2004a).
- Zhang P, Li WL, Fellers J, Friebe B, Gill BS: BAC-FISH in wheat identifies chromosome landmarks consisting of different types of transposable elements. Chromosoma 112:288–299 (2004b).
- Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214 (2000).

ARTICLE

doi:10.1038/nature11543

A physical, genetic and functional sequence assembly of the barley genome

The International Barley Genome Sequencing Consortium*

Barley (*Hordeum vulgare* L.) is among the world's earliest domesticated and most important crop plants. It is diploid with a large haploid genome of 5.1 gigabases (Gb). Here we present an integrated and ordered physical, genetic and functional sequence resource that describes the barley gene-space in a structured whole-genome context. We developed a physical map of 4.98 Gb, with more than 3.90 Gb anchored to a high-resolution genetic map. Projecting a deep whole-genome shotgun assembly, complementary DNA and deep RNA sequence data onto this framework supports 79,379 transcript clusters, including 26,159 'high-confidence' genes with homology support from other plant genomes. Abundant alternative splicing, premature termination codons and novel transcriptionally active regions suggest that post-transcriptional processing forms an important regulatory layer. Survey sequences from diverse accessions reveal a landscape of extensive single-nucleotide variation. Our data provide a platform for both genome-assisted research and enabling contemporary crop improvement.

Cultivated barley, derived from its wild progenitor *Hordeum vulgare* ssp. *spontaneum*, is among the world's earliest domesticated crop species[1] and today represents the fourth most abundant cereal in both area and tonnage harvested (http://faostat.fao.org). Approximately three-quarters of global production is used for animal feed, 20% is malted for use in alcoholic and non-alcoholic beverages, and 5% as an ingredient in a range of food products[2]. Barley is widely adapted to diverse environmental conditions and is more stress tolerant than its close relative wheat[3]. As a result, barley remains a major food source in poorer countries[4], maintaining harvestable yields in harsh and marginal environments. In more developed societies it has recently been classified as a true functional food. Barley grain is particularly high in soluble dietary fibre, which significantly reduces the risk of serious human diseases including type II diabetes, cardiovascular disease and colorectal cancers that afflict hundreds of millions of people worldwide[5]. The USA Food and Drug Administration permit a human health claim for cell-wall polysaccharides from barley grain.

As a diploid, inbreeding, temperate crop, barley has traditionally been considered a model for plant genetic research. Large collections of germplasm containing geographically diverse elite varieties, landraces and wild accessions are readily available[6] and undoubtedly contain alleles that could ameliorate the effect of climate change and further enhance dietary fibre in the grain. Enriching its broad natural diversity, extensive characterized mutant collections containing all of the morphological and developmental variation observed in the species have been generated, characterized and meticulously maintained. The major impediment to the exploitation of these resources in fundamental and breeding science has been the absence of a reference genome sequence, or an appropriate enabling alternative. Providing either of these has been the primary research challenge to the global barley community.

In response to this challenge, we present a novel model for delivering the genome resources needed to reinforce the position of barley as a model for the Triticeae, the tribe that includes bread and durum wheats, barley and rye. We introduce the barley genome gene space, which we define as an integrated, multi-layered informational resource that provides access to the majority of barley genes in a highly structured physical and genetic framework. In association with comparative sequence and transcriptome data, the gene space provides a new molecular and cellular insight into the biology of the species, providing a platform to advance gene discovery and genome-assisted crop improvement.

A sequence-enriched barley physical map

We constructed a genome-wide physical map of the barley cultivar (cv.) Morex by high-information-content fingerprinting[7] and contig assembly[8] of 571,000 bacterial artificial chromosome (BAC) clones (~14-fold haploid genome coverage) originating from six independent BAC libraries[9]. After automated assembly and manual curation, the physical map comprised 9,265 BAC contigs with an estimated N50 contig size of 904 kilobases and a cumulative length of 4.98 Gb (Methods, Supplementary Note 2). It is represented by a minimum tiling path (MTP) of 67,000 BAC clones. Given a genome size of 5.1 Gb[10], more than 95% of the barley genome is represented in the physical map, comparing favourably to the 1,036 contigs that represent 80% of the 1 Gb wheat chromosome 3B[11].

We enhanced the physical map by integrating shotgun sequence information from 5,341 gene-containing[12,13] and 937 randomly selected BAC clones (Methods, Supplementary Notes 2 and 3, and Supplementary Table 4), and 304,523 BAC-end sequence (BES) pairs (Supplementary Table 3). These provided 1,136 megabases (Mb) of genomic sequence integrated directly into the physical map (Supplementary Tables 3 and 4). This framework facilitated the incorporation of whole-genome shotgun sequence data and integration of the physical and genetic maps. We generated whole-genome shotgun sequence data from genomic DNA of cv. 'Morex' by short-read Illumina GAIIx technology, using a combination of 300 base pairs (bp) paired-end and 2.5 kb mate-pair libraries, to >50-fold haploid genome coverage (Supplementary Note 3.3). *De novo* assembly resulted in sequence contigs totalling 1.9 Gb. Due to the high proportion of repetitive DNA, a substantial part of the whole-genome shotgun data collapsed into relatively small contigs characterized by exceptionally high read depths. Overall, 376,261 contigs were larger than 1 kb (N50 = 264,958 contigs, N50 length = 1,425 bp). Of these, 112,989

*A list of authors and their affiliations appears at the end of the paper.

RESEARCH ARTICLE

(308 Mb) could be anchored directly to the sequence-enriched physical map by sequence homology.

We implemented a hierarchical approach to further anchor the physical and genetic maps (Methods, Supplementary Note 4). A total of 3,241 genetically mapped gene-based single-nucleotide variants (SNV) and 498,165 sequence-tag genetic markers[14] allowed us to use sequence homology to assign 4,556 sequence-enriched physical map contigs spanning 3.9 Gb to genetic positions along each barley chromosome. An additional 1,881 contigs were assigned to chromosomal bins by sequence homology to chromosome-arm-specific sequence data sets[15] (Supplementary Note 4.4). Thus, 6,437 physical map contigs totalling 4.56 Gb (90% of the genome), were assigned to chromosome arm bins, the majority in linear order. Non-anchored contigs were typically short and lacked genetically informative sequences required for positional assignment.

Consistent with genome sequences of other grass species[16] the pericentromeric and centromeric regions of barley chromosomes exhibit significantly reduced recombination frequency, a feature that compromises exploitation of genetic diversity and negatively impacts genetic studies and plant breeding. Approximately 1.9 Gb or 48% of the genetically anchored physical map (3.9 Gb) was assigned to these regions (Fig. 1 and Supplementary Fig. 11).

Repetitive nature of the barley genome

A characteristic of the barley genome is the abundance of repetitive DNA[17]. We observed that approximately 84% of the genome is comprised of mobile elements or other repeat structures (Supplementary Note 5). The majority (76% in random BACs) of these consists of retrotransposons, 99.6% of which are long terminal repeat (LTR) retrotransposons. The non-LTR retrotransposons contribute only 0.31% and the DNA transposons 6.3% of the random BAC sequence. In the fraction of the genome with a high proportion of repetitive elements, the LTR *Gypsy* retrotransposon superfamily was 1.5-fold more abundant than the *Copia* superfamily, in contrast to observations in both *Brachypodium*[18] and rice[19]. However, gene-bearing BACs were slightly depleted of retrotransposons, consistent with *Brachypodium*[18] where young *Copia* retroelements are preferentially found in gene-rich, recombinogenic regions from which inactive *Gypsy* retroelements have been lost by LTR–LTR recombination. Overall, we see reduced repetitive DNA content within the terminal 10% of the physical map of each barley chromosome arm (Fig. 1). Class I and II elements show non-quantitative reverse-image distribution along barley chromosomes (Fig. 1), a feature shared with other grass genomes[16,20] and shown by fluorescence *in situ* hybridization (FISH) mapping[17]. Not surprisingly, the whole-genome shotgun assembly shows a lower abundance of LTR retrotransposons (average 53%) than gene-bearing BACs. That LTR retrotransposons are long (~10 kb), highly repetitive and often nested[21] supports our assumption that short reads either collapsed or did not assemble. Short interspersed elements (SINEs)[22], short (80–600 bp) non-autonomous retrotransposons that are highly repeated in barley, showed no differential exclusion from the assemblies. However, miniature inverted-repeat transposable elements (MITEs), small non-autonomous DNA transposons[23], were twofold enriched in the whole-genome shotgun assemblies compared with BES reads or random BACs, consistent with the gene richness of the assemblies and their association with genes[23]. Both MITEs and SINEs are 1.5 to 2-fold enriched in gene-bearing BACs which could indicate that SINEs are also preferentially integrated into gene-rich regions, or because they are older than LTR retroelements, may simply remain visible in and around genes where retro insertions have been selected against.

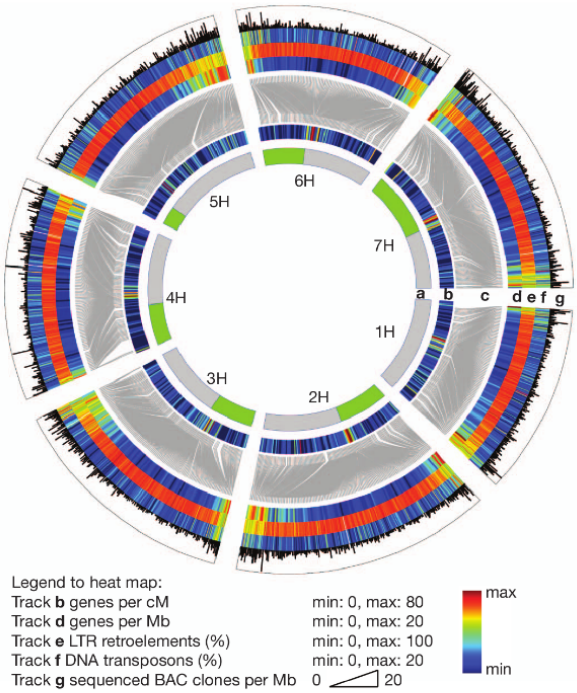

Figure 1 | Landscape of the barley gene space. Track **a** gives the seven barley chromosomes. Green/grey colour depicts the agreement of anchored fingerprint (FPC) contigs with their chromosome arm assignment based on chromosome-arm-specific shotgun sequence reads (for further details see Supplementary Note 4). For 1H only whole-chromosome sequence assignment was available. Track **b**, distribution of high-confidence genes along the genetic map; track **c**, connectors relate gene positions between genetic and the integrated physical map given in track **d**. Position and distribution of track **e** class I LTR-retroelements and track **f** class II DNA transposons are given. Track **g**, distribution and positioning of sequenced BACs.

Transcribed portion of the barley genome

The transcribed complement of the barley gene space was annotated by mapping 1.67 billion RNA-seq reads (167 Gb) obtained from eight stages of barley development as well as 28,592 barley full-length cDNAs[24] to the whole-genome shotgun assembly (Methods, Supplementary Notes 6, 7 and Supplementary Tables 20–22). Exon detection and consensus gene modelling revealed 79,379 transcript clusters, of which 75,258 (95%) were anchored to the whole-genome shotgun assembly (Supplementary Notes 7.1.1 and 7.1.2). Based on a gene-family-directed comparison with the genomes of *Sorghum*, rice, *Brachypodium* and *Arabidopsis*, 26,159 of these transcribed loci fall into clusters and have homology support to at least one reference genome (Supplementary Fig. 16); they were defined as high-confidence genes. Comparison against a data set of metabolic genes in *Arabidopsis thaliana*[25] indicated a detection rate of 86%, allowing the barley gene-set to be estimated as approximately 30,400 genes. Due to lack of homology and missing support from gene family clustering, 53,220 transcript loci were considered low-confidence (Table 1). High-confidence and low-confidence barley genes exhibited distinct characteristics: 75% of the high-confidence genes had a multi-exon structure, compared with only 27% of low-confidence genes (Table 1). The mean size of high-confidence genes was 3,013 bp compared with 972 bp for low-confidence genes. A total of 14,481 low-confidence genes showed distant homology to plant proteins in public databases (Supplementary Notes 7.1.2, 7.1.4 and Supplementary Fig. 18), identifying them as potential gene fragments known to populate Triticeae genomes at high copy number and that often result from transposable element activity[26].

A total of 15,719 high-confidence genes could be directly associated with the genetically anchored physical map (Supplementary Note 4). An additional 3,743 were integrated by invoking a conservation of synteny model (Supplementary Note 4.5) and a further 4,692 by association

Table 1 | Characteristics of high-confidence and low-confidence gene sets in barley

	High confidence	Low confidence
Number of genes	26,159	53,220
Gene loci positioned on barley cultivar Morex assembly*	24,243 (93%)	51,015 (96%)
Single exon	5,954 (25%)	37,395 (73%)
Multi exon	18,289 (75%)	13,620 (27%)
Number of distinct exons†	184,710	107,768
Mean number of distinct exons per gene	7.62	2.11
Number of genes with alternative transcript variants	13,299 (55%)	8,214 (16%)
Total number of predicted transcripts	62,426	69,266
Mean number of transcripts per gene	2.58	1.36
Mean gene locus size (first to last exon)	3,013 bp	972 bp
Mean transcript size (UTR, CDS)	1,878bp	931 bp
Mean exon size	454 bp	536 bp
Gene loci not positioned on barley cv. Morex assembly‡	1,916 (7%)	2,205 (4%)
Tagged by unmapped RNA-seq reads	1,657 (86%)	1,127 (51%)
Not tagged by unmapped RNA-seq reads	259 (14%)	1,078 (49%)

* Gene locus representatives are (1) RNA-seq based transcript or (2) barley fl-cDNA that were mapped to the barley cultivar Morex assembly or tagged by RNA-seq based transcript during clustering.
† Exons of two or more transcripts were counted once if they have identical start and stop positions.
‡ Gene locus representatives are barley fl-cDNAs that were not mapped to the barley cultivar Morex assembly and not matched by any RNA-seq based transcript
CDS, coding sequence.

with chromosome arm whole-genome shotgun data (Supplementary Note 4.4 and Supplementary Table 15). Importantly, the N50 length of whole-genome shotgun sequence contigs containing high-confidence genes was 8,172 bp, which is generally sufficient to include the entire coding sequence, and 5′ and 3′ untranslated regions (UTRs). Overall 24,154 high-confidence genes (92.3%) were associated and positioned in the physical/genetic scaffold, representing a gene density of five genes per Mb. Proximal and distal ends of chromosomes are more gene-rich, on average containing 13 genes per Mb (Fig. 1).

In comparison with sequenced model plant genomes, gene family analysis (Supplementary Note 7.1.3) revealed some gene families that exhibited barley-specific expansion. We defined the functions of members of these families using gene ontology (GO) and PFAM protein motifs (Supplementary Table 25). Gene families with highly overrepresented GO/PFAM terms included genes encoding (1,3)-β-glucan synthases, protease inhibitors, sugar-binding proteins and sugar transporters. NB-ARC (a nucleotide-binding adaptor shared by APAF-1, certain R gene products and CED-4[27]) domain proteins, known to be involved in defence responses, were also overrepresented, including 191 NBS-LRR type genes. These tended to cluster towards the distal regions of barley chromosomes (Supplementary Fig. 17), including a major group on barley chromosome 1HS, co-localizing with the *MLA* powdery mildew resistance gene cluster[28]. Biased allocation to recombination-rich regions provides the genomic environment for generating sequence diversity required to cope with dynamic pathogen populations[29,30]. It is noteworthy that the highly over-represented (1,3)-β-glucan synthase genes have also been implicated in plant–pathogen interactions[31].

Regulation of gene expression

Deep RNA sequence data (RNA-seq) provided insights into the spatial and temporal regulation of gene expression (Supplementary Note 7.2). We found 72–84% of high-confidence genes to be expressed in all spatiotemporal RNA-seq samples (Fig. 2a), slightly lower than reported for rice[32] where ~95% of transcripts were found in more than one developmental or tissue sample. More importantly, 36–55% of high-confidence barley genes seemed to be differentially regulated between samples (Fig. 2b), highlighting the inherent dynamics of barley gene expression.

Two notable features support the importance of post-transcriptional processing as a central regulatory layer (Supplementary Notes 7.3 and 7.4). First, we observed evidence for extensive alternative splicing. Of the intron-containing high-confidence barley genes, 73% had evidence of alternative splicing (55% of the entire high-confidence set). The spatial and temporal distribution of alternative splicing transcripts deviated significantly from the general occurrence of transcripts in the different tissues analysed (Fig. 2c). Only 17% of alternative splicing transcripts were shared among all samples, and 17–27% of the alternative splicing transcripts were detected only in individual samples, indicating pronounced alternative splicing regulation. We found 2,466 premature termination codon-containing (PTC+) alternative splicing transcripts (9.4% of high-confidence genes) (Fig. 2d and Table 2), similar to the percentage of nonsense-mediated decay (NMD)-controlled genes in a wide range of species[33,34]. Premature termination codons activate the NMD pathway[35], which leads to rapid degradation of PTC+ transcripts, and have been associated with transcriptional regulation during disease and stress response in human and *Arabidopsis*, respectively[34,36–39]. The distribution of PTC+ transcripts was strikingly dissimilar, both spatially and temporally, with only 7.4% shared and between 31% and 40% exclusively observed in only a single sample (Fig. 2d). Genes encoding PTC+-containing transcripts show a broad spectrum of GO terms and PFAM domains and are more prevalent in expanded gene families. These observations support a central role for alternative splicing/NMD-dependent decay of PTC+ transcripts as a mechanism that controls the expression of many different barley genes.

Second, recent reports have highlighted the abundance of novel transcriptionally active regions in rice that lack homology to protein-coding genes or open reading frames (ORFs)[40]. In barley as many as 27,009 preferentially single-exon low-confidence genes can be classified as putative novel transcriptionally active regions (Supplementary Note 7.1.4). We investigated their potential significance by comparing the homology of barley novel transcriptionally active regions with the rice and *Brachypodium* genomes that respectively represent 50 and 30 million years of evolutionary divergence[18]. A total of 4,830 and 2,450 novel transcriptionally active regions yielded a homology match to the *Brachypodium* and rice genomes, respectively (intersection of 2,046; BLAST P value $\leq 10^{-5}$), indicating a putative functional role in pre-mRNA processing or other RNA regulatory processes[41,42].

Natural diversity

Barley was domesticated approximately 10,000 years ago[1]. Extensive genotypic analysis of diverse germplasm has revealed that restricted outcrossing (0–1.8%)[43], combined with low recombination in pericentromeric regions, has resulted in modern germplasm that shows limited regional haplotype diversity[44]. We investigated the frequency and distribution of genome diversity by survey sequencing four diverse barley cultivars ('Bowman', 'Barke', 'Igri' and 'Haruna Nijo') and an *H. spontaneum* accession (Methods and Supplementary Note 8) to a depth of 5–25-fold coverage, and mapping sequence reads against the barley cultivar 'Morex' gene space. We identified more than 15 million non-redundant single-nucleotide variants (SNVs). *H. spontaneum* contributed almost twofold more SNV than each of the cultivars (Supplementary Table 28). Up to 6 million SNV per accession could be assigned to chromosome arms, including up to 350,000 associated with exons (Supplementary Table 29). Approximately 50% of the exon-located SNV were integrated into the genetic/physical framework (Fig. 3, Supplementary Table 30 and Supplementary Fig. 31), providing a platform to establish true genome-wide marker technology for high-resolution genetics and genome-assisted breeding.

We observed a decrease in SNV frequency towards the centromeric and peri-centromeric regions of all barley chromosomes, a pattern that seemed more pronounced in the barley cultivars. This trend was supported by SNV identified in RNA-seq data from six additional cultivars mapped onto the Morex genomic assembly (Supplementary Note 8.2). We attribute this pattern of eroded genetic diversity to low recombination in the pericentromeric regions, which reduces effective population size and consequently haplotype diversity. Whereas

RESEARCH ARTICLE

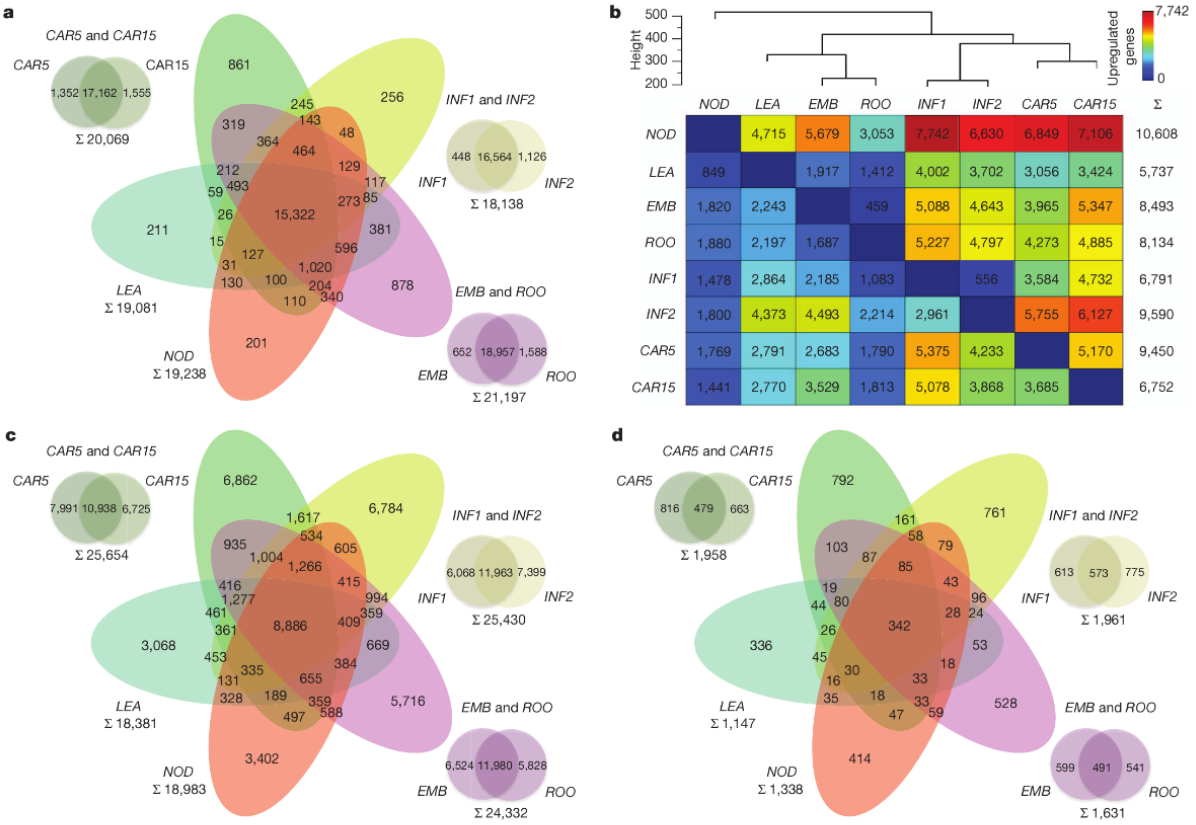

Figure 2 | Atlas of barley gene expression. **a**, Barley gene expression in different spatial and temporal RNA-seq samples (Supplementary Notes 6, 7). Numbers refer to high-confidence genes. **b**, Dendrogram depicting relatedness of samples and colour-coded matrix showing number of significantly upregulated high-confidence genes in pairwise comparisons. Σ, total number of non-redundant high-confidence genes upregulated in comparison to all other samples. Height, complete linkage cluster distance ($\log_2$(fragments per kilobase of exon per million fragments mapped)); see Supplementary Note 7.2.5.1. **c**, Distribution and overlap of alternatively spliced barley transcripts between RNA-seq samples. **d**, Distribution and overlap of alternative splicing transcripts fulfilling criteria for PTC+ as detected in different spatial and temporal RNA-seq samples (Supplementary Note 7.4).

Table 2 | Alternative splicing and transcripts containing PTCs in high-confidence genes

General statistics of alternative splicing in high-confidence genes	
High-confidence genes with RNA-seq data to monitor alternative splicing	24,243
Predicted transcripts at high-confidence genes	62,426
Transcripts with complete CDS structures*	62,256
Transcripts with partial CDS structures†	170
Genes with alternative transcripts	13,299
Predicted transcripts derived from genes with alternative splicing	51,482
Premature stop codon analysis	
Predicted transcripts used for PTC analysis‡	51,338
Transcripts without PTC	41,461 (81%)
Transcripts containing PTC	9,877
PTC caused by intron retention	5,286 (10%)
PTC+ transcripts predicted to be NMD****-sensitive	4,591 (9%)
Gene loci incorporating PTC+/NMD transcripts	2,466

*Entire predicted coding sequence (100%) was transferred to transcript model on barley cultivar Morex contigs.
†Predicted coding sequence could not be completely projected to genomic transcript model (partial mapping of fl-cDNA).
‡Only transcripts with structures for entire coding sequence on barley cultivar Morex WGS assembly were considered.
CDS, coding sequence.

H. spontaneum may serve here as a reservoir of genetic diversity, using this diversity may itself be compromised by restricted recombination and the consequent inability to disrupt tight linkages between desirable and deleterious alleles. Surprisingly, the short arm of chromosome 4H had a significantly lower SNV frequency than all other barley chromosomes (Supplementary Fig. 33). This may be a consequence of a further reduction in recombination frequency on this chromosome, which is genetically (but not physically) shortest. Reduced SNV diversity was also observed in regions we interpret to be either the consequences of recent breeding history or could indicate landmarks of domestication (Fig. 3).

Discussion

The size of Triticeae cereal genomes, due to their highly repetitive DNA composition, has severely compromised the assembly of whole-genome shotgun sequences and formed a barrier to the generation of high-quality reference genomes. We circumvented these problems by integrating complementary and heterogeneous sequence-based genomic and genetic data sets. This involved coupling a deep physical map with high density genetic maps, superimposing deep short-read whole-genome shotgun assemblies, and annotating the resulting linear, albeit punctuated, genomic sequence with deep-coverage RNA-derived data (full-length cDNA and RNA-seq). This allowed us to systematically delineate approximately 4 Gb (80%) of the genome, including more than 90% of the expressed genes. The resulting genomic framework

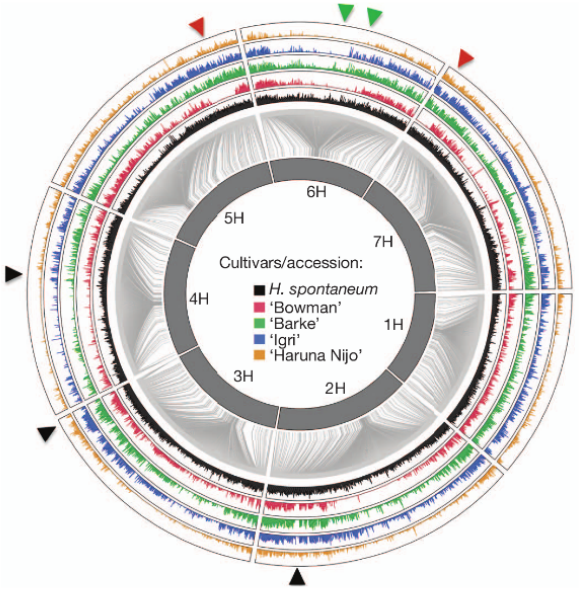

Figure 3 | Single nucleotide variation (SNV) frequency in barley. Barley chromosomes indicated as inner circle of grey bars. Connector lines give the genetic/physical relationship in the barley genome. SNV frequency distribution displayed as five coloured circular histograms (scale, relative abundance of SNVs within accession; abundance, total number of SNVs in non-overlapping 50-kb intervals of concatenated 'Morex' genomic scaffold; range, zero to maximum number of SNVs per 50-kb interval). Selected patterns of SNV frequency indicated by coloured arrowheads (for further details see Supplementary Note 8). Colouring of arrowheads refers to cultivar with deviating SNV frequency for the respective region.

provides a detailed insight into the physical distribution of genes and repetitive DNA and how these features relate to genetic characteristics such as recombination frequency, gene expression and patterns of genetic variation.

The centromeric and peri-centromeric regions of barley chromosomes contain a large number of functional genes that are locked into recombinationally 'inert' genomic regions[15,16]. The gene-space distribution highlights that these regions expand to almost 50% of the physical length of individual chromosomes. Given well-established levels of conserved synteny, this will probably be a general feature of related grass genomes that will have important practical implications. For example, infrequent recombination could function to maintain evolutionarily selected and co-adapted gene complexes. It will certainly restrict the release of the genetic diversity required to decouple advantageous from deleterious alleles, a potential key to improving genetic gain. Understanding these effects will have important consequences for crop improvement. Moreover, for gene discovery, forward genetic strategies based on recombination will not be effective in these regions. Whereas alternative approaches exist for some targets (for example, by coupling resequencing technologies with collections of natural or induced mutant alleles), for most traits it remains a serious impediment. Some promise may lie in manipulating patterns of recombination by either genetic or environmental intervention[47]. Quite strikingly, our data also reveal that a complex layer of post-transcriptional regulation will need to be considered when attempting to link barley genes to functions. Connections between post-transcriptional regulation such as alternative splicing and functional biological consequences remain limited to a few specific examples[48], but the scale of our observations suggest this list will expand considerably.

In conclusion, the barley gene space reported here provides an essential reference for genetic research and breeding. It represents a hub for trait isolation, understanding and exploiting natural genetic diversity and investigating the unique biology and evolution of one of the world's first domesticated crops.

METHODS SUMMARY

Methods are available in the online version of the paper.

Full Methods and any associated references are available in the online version of the paper.

Received 2 May; accepted 30 August 2012.
Published online 17 October 2012.

1. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457,** 843–848 (2009).
2. Blake, T., Blake, V., Bowman, J. & Abdel-Haleem, H. in *Barley: Production, Improvement and Uses* (ed. S. E. Ullrich) 522–531 (Wiley-Blackwell, 2011).
3. Nevo, E. *et al.* Evolution of wild cereals during 28 years of global warming in Israel. *Proc. Natl Acad. Sci. USA* **109,** 3412–3415 (2012).
4. Grando, S. & Macpherson, H. G. in *Proceedings of the International Workshop on Food Barley Improvement,* 14–17 January 2002, Hammamet, Tunisia 156 (ICARDA, Aleppo, Syria, 2005).
5. Collins, H. M. *et al.* Variability in fine structures of noncellulosic cell wall polysaccharides from cereal grains: potential importance in human health and nutrition. *Cereal Chem.* **87,** 272–282 (2010).
6. Bockelman, H. E. & Valkoun, J. in *Barley: Production, Improvement, and Uses* (ed. S. E. Ullrich) 144–159 (Wiley-Blackwell, 2011).
7. Luo, M.-C. *et al.* High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82,** 378–389 (2003).
8. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs built with fingerprints, markers, and FPC V4.7. *Genome Res.* **10,** 1772–1787 (2000).
9. Schulte, D. *et al.* BAC library resources for map-based cloning and physical map construction in barley (*Hordeum vulgare* L.). *BMC Genomics* **12,** 247 (2011).
10. Doležel, J. *et al.* Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.* **82,** 17–26 (1998).
11. Paux, E. *et al.* A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322,** 101–104 (2008).
12. Madishetty, K., Condamine, P., Svensson, J. T., Rodriguez, E. & Close, T. J. An improved method to identify BAC clones using pooled overgos. *Nucleic Acids Res.* **35,** e5 (2007).
13. Lonardi, S. *et al.* Barcoding-free BAC pooling enables combinatorial selective sequencing of the barley gene space. preprint at http://arxiv.org/abs/1112.4438 (2011).
14. Poland, J. A., Brown, P. J., Sorrells, M. E. & Jannink, J.-L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **7,** e32253 (2012).
15. Mayer, K. F. X. *et al.* Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23,** 1249–1263 (2011).
16. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326,** 1112–1115 (2009).
17. Wicker, T. *et al.* A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.* **59,** 712–722 (2009).
18. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon. Nature* **463,** 763–768 (2010).
19. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
20. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).
21. Kronmiller, B. A. & Wise, R. P. TEnest: automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol.* **146,** 45–59 (2008).
22. Ohshima, K. & Okada, N. SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet. Genome Res.* **110,** 475–490 (2005).
23. Wessler, S. R., Bureau, T. & White, S. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5,** 814–821 (1995).
24. Matsumoto, T. *et al.* Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* **156,** 20–28 (2011).
25. Zhang, P. *et al.* MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiol.* **138,** 27–37 (2005).
26. Wicker, T. *et al.* Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* **23,** 1706–1718 (2011).
27. van der Biezen, E. A. & Jones, J. D. G. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8,** R226–R228 (1998).
28. Wei, F., Wing, R. A. & Wise, R. P. Genome dynamics and evolution of the *Mla* (powdery mildew) resistance locus in barley. *Plant Cell* **14,** 1903–1917 (2002).
29. Halterman, D. A. & Wise, R. P. A single-amino acid substitution in the sixth leucine-rich repeat of barley *MLA6* and *MLA13* alleviates dependence on *RAR1* for disease resistance signaling. *Plant J.* **38,** 215–226 (2004).

RESEARCH ARTICLE

30. Seeholzer, S. *et al.* Diversity at the *Mla* powdery mildew resistance locus from cultivated barley reveals sites of positive selection. *Mol. Plant Microbe Interact.* **23,** 497–509 (2010).
31. Jacobs, A. K. *et al.* An *Arabidopsis* callose synthase, GSL5, is required for wound and papillary callose formation. *Plant Cell* **15,** 2503–2513 (2003).
32. Jiao, Y. *et al.* A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nature Genet.* **41,** 258–263 (2009).
33. Conti, E. & Izaurralde, E. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.* **17,** 316–325 (2005).
34. Kalyna, M. *et al.* Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in *Arabidopsis*. *Nucleic Acids Res.* **40,** 2454–2469 (2012).
35. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100,** 189–192 (2003).
36. Bhuvanagiri, M., Schlitter, A. M., Hentze, M. W. & Kulozik, A. E. NMD: RNA biology meets human genetic medicine. *Biochem. J.* **430,** 365–377 (2010).
37. Rayson, S. *et al.* A role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in *Arabidopsis thaliana* NMD mutants. *PLoS ONE* **7,** e31917 (2012).
38. Riehs-Kearnan, N., Gloggnitzer, J., Dekrout, B., Jonak, C. & Riha, K. Aberrant growth and lethality of *Arabidopsis* deficient in nonsense-mediated RNA decay factors is caused by autoimmune-like response. *Nucleic Acids Res.* **40,** 5615–5624 (2012).
39. Jeong, H.-J. *et al.* Nonsense-mediated mRNA decay factors, *UPF1* and *UPF3*, contribute to plant defense. *Plant Cell Physiol.* **52,** 2147–2156 (2011).
40. Lu, T. *et al.* Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20,** 1238–1249 (2010).
41. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482,** 339–346 (2012).
42. Chinen, M. & Tani, T. Diverse functions of nuclear non-coding RNAs in eukaryotic gene expression. *Front. Biosci.* **17,** 1402–1417 (2012).
43. Abdel-Ghani, A. H., Parzies, H. K., Omary, A. & Geiger, H. H. Estimating the outcrossing rate of barley landraces and wild barley populations collected from ecologically different regions of Jordan. *Theor. Appl. Genet.* **109,** 588–595 (2004).
44. Comadran, J. *et al.* Patterns of polymorphism and linkage disequilibrium in cultivated barley. *Theor. Appl. Genet.* **122,** 523–531 (2011).
45. Close, T. J. *et al.* Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10,** 582 (2009).
46. Thiel, T. *et al.* Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol. Biol.* **9,** 209 (2009).
47. Martinez-Perez, E. & Moore, G. To check or not to check? The application of meiotic studies to plant breeding. *Curr. Opin. Plant Biol.* **11,** 222–227 (2008).
48. Halterman, D. A., Wei, F. S. & Wise, R. P. Powdery mildew-induced *Mla* mRNAs are alternatively spliced and contain multiple upstream open reading frames. *Plant Physiol.* **131,** 558–567 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work has been supported from the following funding sources: German Ministry of Education and Research (BMBF) grant 0314000 "BARLEX" to K.F.X.M., M.P., U.S. and N.S.; Leibniz Association grant (Pakt f. Forschung und Innovation) to N.S.; European project of the 7th framework programme "TriticeaeGenome" to R.W., A.S., K.F.X.M., M.M. and N.S.; SFB F3705, of the Austrian Wissenschaftsfond (FWF) to K.F.X.M.; ERA-NET PG project "BARCODE" grant to M.M., N.S. and R.W.; Scottish Government/BBSRC grant BB/100663X/1 to R.W., D.M., P.H., J.R., M.C. and P.K.; National Science Foundation grant DBI 0321756 "Coupling EST and Bacterial Artificial Chromosome Resources to Access the Barley Genome" and DBI-1062301 "Barcoding-Free Multiplexing: Leveraging Combinatorial Pooling for High-Throughput Sequencing" to T.J.C. and S.L.; USDA-CSREES-NRI grant 2006-55606-16722 "Barley Coordinated Agricultural Project: Leveraging Genomics, Genetics, and Breeding for Gene Discovery and Barley Improvement" to G.J.M., R.P.W., T.J.C. and S.L.; the Agriculture and Food Research Initiative Plant Genome, Genetics and Breeding Program of USDA-CSREES-NIFA grant 2009-65300-05645 "Advancing the Barley Genome" to T.J.C., S.L. and G.J.M.; BRAIN and NBRP-Japan grants to K.S., Japanese MAFF Grant (TRG1008) to T.M. A full list of acknowledgements is in the Supplementary Information.

Author Contributions See list of consortium authors. R.A., D.S., H.L., B.S., S.T., M.G., F.C., T.N., M.S., M.P., H.G., P.H., T.S., K.F.X.M., R.W. and N.S. contributed equally to their respective work packages and tasks.

Author Information Sequence resources generated or compiled in this study have been deposited at EMBL/ENA or NCBI GenBank. A full list of sequence raw data accession numbers as well as URLs for data download, visualization or search are provided in Supplementary Note 1 and Supplementary Table 1. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike license, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.F.X.M (k.mayer@helmholtz-muenchen.de), R.W. (Robbie.Waugh@hutton.ac.uk) or N.S. (stein@ipk-gatersleben.de).

The International Barley Genome Sequencing Consortium (IBSC)

Principal investigators Klaus F. X. Mayer[1], Robbie Waugh[2], Peter Langridge[3], Timothy J. Close[4], Roger P. Wise[5], Andreas Graner[6], Takashi Matsumoto[7], Kazuhiro Sato[8], Alan Schulman[9], Gary J. Muehlbauer[10], Nils Stein[6]

Physical map construction and direct anchoring Ruvini Ariyadasa[6], Daniela Schulte[6], Naser Poursarebani[6], Ruonan Zhou[6], Burkhard Steuernagel[6], Martin Mascher[6], Uwe Scholz[6], Bujun Shi[3], Peter Langridge[3], Kavitha Madishetty[4], Jan T. Svensson[4], Prasanna Bhat[4], Matthew Moscou[4], Josh Resnik[4], Timothy J. Close[4], Gary J. Muehlbauer[10], Pete Hedley[2], Hui Liu[2], Jenny Morris[2], Robbie Waugh[2], Zeev Frenkel[11], Avraham Korol[11], Hélène Bergès[12], Andreas Graner[6], Nils Stein (leader)[6]

Genomic sequencing and assembly Burkhard Steuernagel[6], Uwe Scholz[6], Stefan Taudien[13], Marius Felder[13], Marco Groth[13], Matthias Platzer[13], Nils Stein (leader)[6]

BAC sequencing and assembly Burkhard Steuernagel[6], Uwe Scholz[6], Axel Himmelbach[6], Stefan Taudien[13], Marius Felder[13], Matthias Platzer[13], Stefano Lonardi[14], Denisa Duma[14], Matthew Alpert[14], Francesa Cordero[14,22], Marco Beccuti[14], Gianfranco Ciardo[14], Yaqin Ma[14], Steve Wanamaker[4], Timothy J. Close (co-leader)[4], Nils Stein (leader)[6]

BAC-end sequencing Federica Cattonaro[15], Vera Vendramin[16], Simone Scalabrin[15], Slobodanka Radovic[16], Rod Wing[17], Daniela Schulte[6], Burkhard Steuernagel[6], Michele Morgante[15,16], Nils Stein[6], Robbie Waugh (leader)[2]

Integration of physical/genetic map and sequence resources Thomas Nussbaumer[1], Heidrun Gundlach[1], Mihaela Martis[1], Ruvini Ariyadasa[6], Naser Poursarebani[6], Burkhard Steuernagel[6], Uwe Scholz[6], Roger P. Wise[5], Jesse Poland[18], Nils Stein[6], Klaus F. X. Mayer (leader)[1]

Gene annotation Manuel Spannagl[1], Matthias Pfeifer[1], Heidrun Gundlach[1], Klaus F. X. Mayer (leader)[1]

Repetitive DNA analysis Heidrun Gundlach[1], Cédric Moisy[9], Jaakko Tanskanen[9], Simone Scalabrin[15], Andrea Zuccolo[15], Vera Vendramin[16], Michele Morgante[15,16], Klaus F. X. Mayer (co-leader)[1], Alan Schulman (leader)[9]

Transcriptome sequencing and analysis Matthias Pfeifer[1], Manuel Spannagl[1], Pete Hedley[2], Jenny Morris[2], Joanne Russell[2], Arnis Druka[2], David Marshall[2], Micha Bayer[2], David Swarbreck[19], Dharanya Sampath[19], Sarah Ayling[19], Melanie Febrer[19], Mario Caccamo[19], Takashi Matsumoto[7], Tsuyoshi Tanaka[7], Kazuhiro Sato[8], Roger P. Wise[5], Timothy J. Close[4], Steve Wannamaker[4], Gary J. Muehlbauer[10], Nils Stein[6], Klaus F. X. Mayer (co-leader)[1], Robbie Waugh (leader)[2]

Re-sequencing and diversity analysis Burkhard Steuernagel[6], Thomas Schmutzer[6], Martin Mascher[6], Uwe Scholz[6], Stefan Taudien[13], Matthias Platzer[13], Kazuhiro Sato[8], David Marshall[2], Micha Bayer[2], Robbie Waugh (co-leader)[2], Nils Stein (leader)[6]

Writing and editing of the manuscript Klaus F. X. Mayer (co-leader)[1], Robbie Waugh (co-leader)[2], John W. S. Brown[2,20], Alan Schulman[9], Peter Langridge[3], Matthias Platzer[13], Geoffrey B. Fincher[21], Gary J. Muehlbauer[10], Kazuhiro Sato[8], Timothy J. Close[4], Roger P. Wise[5] & Nils Stein (leader)[6]

[1]MIPS/IBIS, Helmholtz Zentrum München, D-85764 Neuherberg, Germany. [2]The James Hutton Institute, Invergowrie, Dundee DD2 5DE, UK. [3]Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond 5064, Australia. [4]Department of Botany & Plant Sciences, University of California, Riverside, California 92521, USA. [5]USDA-ARS, Department of Plant Pathology & Microbiology, Iowa State University, Ames, Iowa 50011-1020, USA. [6]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Seeland OT Gatersleben, Germany. [7]National Institute of Agrobiological Sciences, 2-1-2, Kannondai, Tsukuba Ibaraki 305-8602, Japan. [8]Okayama University, Kurashiki 710-0046, Japan. [9]MTT Agrifood Research and Institute of Biotechnology, University of Helsinki, FIN-00014 Helsinki, Finland. [10]University of Minnesota, Department of Agronomy and Plant Genetics, Department of Plant Biology, St Paul, Minnesota 55108, USA. [11]Institute of Evolution, University of Haifa, Haifa 31905, Israel. [12]INRA-CNRGV, Auzeville CS 52627, France. [13]Leibniz Institute of Age Research- Fritz Lipmann Institute (FLI), D-07745 Jena, Germany. [14]Department of Computer Science & Engineering, University of California, Riverside, California 92521, USA. [15]Istituto di Genomica Applicata, Via J. Linussio 51, 33100 Udine, Italy. [16]Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, 33100 Udine, Italy. [17]University of Arizona, Arizona Genomics Institute, Tucson, Arizona 85721, USA. [18]USDA-ARS Hard Winter Wheat Genetics Research Unit and Kansas State University, Manhattan, Kansas 66506, USA. [19]The Genome Analysis Centre, Norwich Research Park, Norwich NR4 7UH, UK. [20]Division of Plant Sciences, University of Dundee at The James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK. [21]ARC Centre of Excellence in Plant Cell Walls, University of Adelaide, Waite Campus, Glen Osmond, South Australia 5064, Australia. [22]Department of Computer Science, Corso Svizzera 185, 10149 Torino, Italy.

METHODS

Building the physical map. BAC clones of six libraries of cultivar 'Morex'[9,49] were analysed by high information content fingerprinting (HICF)[7,9]. A total of 571,000 edited profiles was assembled using FPC v9.2[8] (Supplementary Table 2) (Sulston score threshold of 10^{-90}, tolerance = 5, tolerated Q clones = 10%). Nine iterative automated re-assemblies were performed at successively reduced stringency (Sulston score of 10^{-85} to 10^{-45}). A final step of manual merging of FPC contigs was performed at lower stringency (Sulston score threshold 10^{-25}) considering genetic anchoring information for markers with a genetic distance $\leq \pm 5$ cM. This produced 9,265 FPcontigs (approximately 14-fold haploid genome coverage) (Supplementary Table 2).

Genomic sequencing. BAC-end sequencing (BES). BAC insert ends were sequenced using Sanger sequencing (Supplementary Note 2.1). Vector and quality trimming of sequence trace files was conducted using LUCY[50] (http://www.jcvi.org/cms/research/software/). Short reads (that is, < 100 bp) were removed. Organellar DNA and barley pathogen sequences were filtered by BLASTN comparisons to public sequence databases (http://www.ncbi.nlm.nih.gov/).

BAC shotgun sequencing (BACseq). Seed BACs of the FPC map were sequenced to reveal gene sequence information for physical map anchoring. 4,095 BAC clones were shotgun sequenced in pools of 2 × 48 individually barcoded BACs on Roche/454 GS FLX or FLX Titanium[51,52]. Sequences were assembled using MIRA v3.2.0 (http://www.chevreux.org/projects_mira.html) at default parameters with features 'accurate', '454', 'genome', 'denovo'. An additional 2,183 gene-bearing BACs (Supplementary Note 3.2) were sequenced using Illumina HiSeq 2000 in 91 combinatorial pools[13]. Deconvoluted reads were assembled using VELVET[53]. Assembly statistics are given in Supplementary Table 4.

Whole-genome shotgun sequencing. Illumina paired-end (PE; fragment size ~350 bp) and mate-pair (MP; fragment size ~2.5 kb) libraries were generated from fragmented genomic DNA[54] of different barley cultivars ('Morex', 'Barke', 'Bowman', 'Igri') and an S3 single-seed selection of a wild barley accession B1K-04-12[55] (*Hordeum vulgare* ssp. *spontaneum*). Libraries were sequenced by Illumina GAIIx and Hiseq 2000. Genomic DNA of cultivar 'Haruna Nijo' (size range of 600–1,000 bp) was sequenced using Roche 454 GSFLX Titanium chemistry.

Whole-genome shotgun sequence assembly. PE and MP whole-genome shotgun libraries were calibrated for fragment sizes by mapping pairs against the chloroplast sequence of barley (NC_008590) using BWA[56]. Sequences were quality trimmed and *de novo* assembled using CLC Assembly Cell v3.2.2 (http://www.clcbio.com/). Independent *de novo* assemblies were performed from data of cultivars 'Morex', 'Bowman' and 'Barke'.

Transcriptome sequencing. Eight tissues of cultivar 'Morex' (three biological replications each) earmarking stages of the barley life cycle from germinating grain to maturing caryopsis were selected for deep RNA sequencing (RNA-seq). Plant growth, sampling and sequencing is detailed in Supplementary Information (Supplementary Note 6). Further mRNA sequencing data was generated from eight additional spring barley cultivars within a separate study and was used here for sequence diversity analysis (Supplementary Note 8.2).

Genetic framework of the physical map. The genetic framework for anchoring the physical map of barley was built on a single-nucleotide variation (SNV) map[57] (Supplementary Note 4.3) which provided the highest marker density (3,973) and resolution ($N = 360$, RIL/F8) for a single bi-parental mapping population in barley. Additional high-density genetic marker maps (Supplementary Note 4.3) were compared and aligned on the basis of shared markers. Furthermore, we used genotyping-by-sequencing (GBS)[58] to generate high-density genetic maps comprising 34,396 SNVs and 21,384 SNVs as well as 241,159 and 184,796 dominant (presence/absence) tags for the two doubled haploid populations Oregon Wolfe Barley[14] and Morex × Barke[45], respectively. Altogether 498,165 marker sequence tags were used (Supplementary Table 11).

Genetic anchoring. Genetic integration of the physical map involved procedures of direct and indirect anchoring.

Direct anchoring. Genetic markers were assigned to BAC clones/BAC contigs by three different procedures (Supplementary Note 4.3 and Supplementary Table 9). 2,032 PCR-based markers from published genetic maps[59,60] were PCR-screened on custom multidimensional (MD) DNA pools (http://ampliconexpress.com/) obtained from BAC library HVVMRXALLeA[9]. A single haploid genome equivalent of these MD pools was used for multiplexed screening of 42,302 barley EST-derived unigenes represented on a custom 44K Agilent microarray as previously described[61]. 27,231 barley unigenes, comprising 1,121 with a genetic map position[45,62], could be assigned to 12,313 BACs. 14,600 clones from BAC library HVVMRXALLhA were screened with 3,072 SNP markers on Illumina GoldenGate assays[45] leading to 1,967 markers directly assigned to BACs[13]; approximately one third of this information has been included in the present work.

Indirect anchoring. Sequence resources associated with the FPCmap framework provided the basis for extensive in silico integration of genetic marker information (Supplementary Note 4.3 and Supplementary Table 11). Repeat masked BES sequences, sequences of anchored markers and 6,295 sequenced BACs allowed integration of 307 Mb of 'Morex' whole-genome shotgun contigs into the FPC map. Genetic markers and barley gene sequences were positioned to this reference by strict sequence homology association. Overall 8,170 (~4.6 Gb) BAC contigs received sequence and/or anchoring information (Supplementary Note 4). 4,556 FPC contigs ($\Sigma = 3.9$ Gb) were anchored to the genetic framework.

Analysis of repetitive DNA and repeat masking. Repeat detection and analysis was undertaken as previously described[18,20] with the exception of an updated repeat library complemented by *de novo* detected repetitive elements from barley (Supplementary Note 5).

Gene annotation, functional categorization and differential expression. Publically available barley full-length cDNAs[24] and RNA-seq data generated in the project (Supplementary Note 6) were used for structural gene calling (Supplementary Note 7). Full-length cDNAs and RNA-seq data were anchored to repeat masked whole-genome shotgun sequence contigs using GenomeThreader[63] and CuffLinks[64], respectively, the latter providing also information of alternatively spliced transcripts. Structural gene calls were combined and the longest ORF for each locus was used as representative for gene family analysis (Supplementary Note 7.1.2).

Gene family clustering was undertaken using OrthoMCL (Supplementary Note 7.1.3) by comparing against the genomes of *Oryza sativa* (RAP2), *Sorghum bicolor*, *Brachypodium distachyon* (v 1.4) and *Arabidopsis thaliana* (TAIR10 release).

Analysis of differential gene expression (Supplementary Note 7.2) was performed on RNA-seq data using CuffDiff[65].

Analysis of sequence diversity. Genome-wide SNV was assessed by mapping (BWA v0.5.9-r16[56]) the original sequence reads of sequenced genotypes to a *de novo* assembly of cultivar 'Morex'. Sequence reads from RNA-seq were mapped against the 'Morex' assembly. Details are provided in Supplementary Note 8.

49. Yu, Y. *et al.* A bacterial artificial chromosome library for barley (*Hordeum vulgare* L.) and the identification of clones containing putative resistance genes. *Theor. Appl. Genet.* **101,** 1093–1099 (2000).
50. Chou, H.-H. & Holmes, M. H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17,** 1093–1104 (2001).
51. Steuernagel, B. *et al. De novo* 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* **10,** 547 (2009).
52. Taudien, S. *et al.* Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Res. Notes* **4,** 411 (2011).
53. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
54. Stein, N., Herren, G. & Keller, B. A new DNA extraction method for high-throughput marker analysis in a large-genome species such as *Triticum aestivum*. *Plant Breed.* **120,** 354–356 (2001).
55. Hübner, S. *et al.* Strong correlation of the population structure of wild barley (*Hordeum spontaneum*) across Israel with temperature and precipitation variation. *Mol. Ecol.* **18,** 1523–1536 (2009).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
57. Comadran, J. *et al.* A homologue of *Antirrhinum CENTRORADIALIS* is a component of the quantitative photoperiod and vernalization independent *EARLINESS PER SE 2* locus in cultivated barley. *Nature Genet.* (in the press).
58. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6,** e19379 (2011).
59. Sato, K., Nankaku, N. & Takeda, K. A high-density transcript linkage map of barley derived from a single population. *Heredity* **103,** 110–117 (2009).
60. Stein, N. *et al.* A 1000 loci transcript map of the barley genome – new anchoring points for integrative grass genomics. *Theor. Appl. Genet.* **114,** 823–839 (2007).
61. Liu, H. *et al.* Highly parallel gene-to-BAC addressing using microarrays. *Biotechniques* **50,** 165–174 (2011).
62. Potokina, E. *et al.* Gene expression quantitative trait locus analysis of 16,000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.* **53,** 90–101 (2008).
63. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.* **47,** 965–978 (2005).
64. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (2012).
65. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).

4. Discussion

4.1. Evaluation of DNA sequencing strategies

The advances of next-generation sequencing led to a broad application of molecular genetic methods [171]. Utilization of cost-efficient whole-genome shotgun sequencing or strategies of complexity reduced sequencing enabled the representation of the genetic diversity even in large collections of plant species. In consequence, several new diversity resources have been established to gain insight of the genetic architecture and diversity. These resources for the crops rye, maize, and barley are integrated in this thesis by four publications and will be discussed in the following sections. The massive reduction in sequencing cost paved the way to encode the genetic makeup of plant species. But complex genomes still suffer the drawback of a large, highly repetitive sequence that is limiting a cost-efficient access into the diverse repertoire. In this respect the rye genome (7.9 Gbp) is still a challenge in terms of required financial support as well as computational complexity. Even today with massive sequencing outcomes it would require a minimum of two and a half complete sequencing runs (20 lanes) of an Illumina HiSeq 2000 (with estimated outcome ~200 Gbp per run) to achieve a 50-fold coverage of the rye genome. Mere sequencing depth is an overestimation of a precise genome coverage calculation. As indicated by Schatz et al. [172] only 70 – 75% of the gained sequencing outcome can be aligned to the reference genome, whereas the remaining raw data lacks sufficient read quality. This emphasizes the need for quality correction and removal of low quality data prior to downstream analysis. In consequence this rough estimate to reach a 50-fold coverage in rye still might be an underestimation. To compare, in the model plant species *A. thaliana* with a 120 Mbp genome size, this would mean to fund almost one fifth of the sequencing costs of the 1001 Genomes Project [57]. In other words, the financial engagement of a comparable volume required for establishing a rye reference in *A. thaliana* would provide the opportunity to sequence one reference strain at 50-fold and in addition up to 200 more strains at 17-fold coverage. This comparison shows that projects for plants with highly repetitive genomes need a different strategy, based on complexity reduction.

4.1.1. Established diversity resources

In this section I will briefly outline which diversity resources were built for rye [2.1], maize [2.2], and barley [2.4]. The ordering of the projects and related results reflects the applied genome complexity reduction staring with the largest reduction (RNA-seq) and finishing with no reduction (WGS-seq).

RYE – RNA-seq

The application of 454 sequencing to the transcriptome of diverse genotypes was proven to be a useful tool for the investigation of diversity in plant genomes that lack a genome reference as shown in maize [74]. To reveal the genetic diversity in rye [2.1], a collection of 5 genotypes was sequenced with the 454 GS FLX sequencer, comprising a total of 2,576,822 single end reads (Table 2). To capture most of the gene repertoire expressed in rye as a response to cold and drought stress, a total of 20 different samples per inbred line was obtained from a set of plant tissues that underwent three different stress treatments (cold stress, dehydration shock, and nutrient-starvation). In rye, the average gene length is estimated to be 2,016 bp. The estimation uses 16,407 genes of the highest confidence class (unpublished data). Compared to the estimated mean gene size in barley stated to be 3,013 bp for high-confidence genes [173] the length is lower in rye. This is explained by additional sequence data sets available in barley compromising eight RNA-seq libraries and in addition a full-length cDNA library that was used for gene model prediction. This led to a more evidence-based gene prediction compared to the synteny based approach used in rye. In consequence, the calculation in rye might be an underestimation and therefore the total transcriptome length is estimated to lay between 56 and 84 Mbp, based on estimated values in rye and barley. The variant discovery detected 277,033 candidate polymorphic positions (raw variants) by applying the tool GigaBayes [160]. The final set comprised 17,917 filtered SNVs that were manually inspected to select a total of 5,234 SNVs for the design of a custom SNP assay (Rye5k) for genotyping (Table 2). The suitability of this DNA sequencing method to utilize 454 transcriptome sequencing for variant detection was later successfully applied in other plant species like wheat [130] or pea [174].

MAIZE – CAP-seq

In the maize publication [2.2] the diversity of 21 maize inbred lines including the genome reference line B73 was assessed by targeted sequencing (CAP-seq) using 454 sequencing. With respect to the project perspective 4,648 genes were selected that are considered to have an effect for biomass

accumulation in maize. The method required an initial design of a customized oligo assay and in consequence is associated with an increased complexity, including knowledge about the design process to evaluate and optimize the target assay. In the final design, exons and introns of selected target genes including the 5'- and 3'-untranslated regions (UTR) were captured. The UTR has been shown to have regulative influence on gene expression [175,176]. We therefore exploited the benefit of CAP-seq and extended our diversity study towards SNVs in UTR sequences. Variant discovery was performed using a large collection of read alignment and variant calling tools to optimize the detection process. In the initial discovery, approximately 4.8 million candidate SNVs were identified (Table 2). This non-redundant number was condensed to 696,665 SNVs by basic filtering discarding all SNVs with coverage below four reads. The final set of variants comprised 383,145 SNVs with a global average number of 45,594 per genotype with a range of 11,386 to 96,158 for non-reference genotypes B106 and UH007, respectively. The targeted sequencing approach was efficient for determining variants in selected genes and in addition also allowed for hypotheses of presence/absence variation (PAV) in genes by comparing different inbred lines. Compared to an RNA-seq experiment it is worth noting that such a PAV analysis would not directly allow meaningful conclusions. In RNA-seq, the pure absence of sequence information of a particular gene could be a result of low or no expression [65]. Therefore, it can be concluded that CAP-seq is more robust because it is not affected by expression differences of studied samples.

BARLEY – WGS-seq

Assessing the diversity through WGS sequencing is a strategy applied in various cultivated plants like rice [177] or maize [3]. Although tremendously challenging in terms of cost and computational requirements, the perspective of gaining insights to the genomic architecture and diversity is convincing.

Diversity of six barley cultivars was assessed by WGS sequencing. The four cultivars 'Morex', 'Barke', 'Bowman', and 'Igri' as well as a wild barley of accession *H. vulgare ssp. spontaneum* were sequenced on the Illumina HiSeq2000 platform. In addition, the cultivar 'Haruna Nijo' was sequenced on a 454 GS FLX. WGS sequence data of cultivar 'Morex' was utilized to construct a *de novo* WGS assembly reference. The initial variant calling was performed by a joint application using SAMtools [153] to construct the required pileup format that subsequently was analyzed by VCFtools [144]. Detected variants were quality filtered (basic filtering) to comprise a robust diversity set. In cultivar 'Morex' quality enhancement was conducted by the removal of 12,065,380 positions that were classified as ambiguous (S8.1 supplemental material of IBSC, 2012). This was intended to gain

Table 2 Overview of three DNA sequencing strategies and their application in diversity studies.

	RNA-seq	CAP-seq	WGS-seq
Species	*Secale cereale* L.	*Zea mays* L.	*Hordeum vulgare* L.
Project	RYE-Express	CornFed	IBSC
Year of publication	2011	2015	2012
Number of studied genotypes / cultivars	5	21	6
Target region	transcriptome	target genes	genome
Total size of target region (Mbp)	56[a]	29	5,100
Sequencing technology	454 (GS FLX)	454 (GS FLX+)	Illumina (HiSeq 2000 & GAIIx) + 454 (GS FLX)
Read number	2.5 million	17.8 million	~90 million (SE) + ~2.500 million (PE)
Read type	single end (SE)	single end (SE)	single end (SE) + paired end (PE)
Total sequence length (Gbp)	0.548	6.432	411
Estimated coverage (x-fold) at target region per genotype (min - max)	2.0 (1.3 - 2.6)	10.6 (5.1 - 15.3)	13.4 (6.5 - 35.2)
SNP number (raw)	>277,000	>4.8 million	>34 million
SNP number (final)	17,917	383,145	>15 million
SNP number (in coding region of genes)	17,917	86,875	23,511[b]
Approx. sequencing cost[c]	+	++	+++

[a] Referring to an estimated number of 28,000 rye genes and mean size of high-confidence gene class I (unpublished data).
[b] SNPs located in exons of high confidence genes. Variants fulfill minimal required score value of 50 and do not overlap to an ambiguous reference position.
[c] For simplicity sequencing costs are ranked from low ('+') to high ('+++').

more confidence and implemented by masking polymorphic sites discovered throughout an alignment of reads of cultivar 'Morex' against the established 'Morex' reference. The complex approach was necessary due to flaws of the WGS reference assembly like the high grade of fragmentation. This limitation is often observed in WGS assemblies and it hinders analysis especially in complex regions corresponding to nested repeats [178]. The majority of these discarded positions are located in short contigs and/or intensified towards the end of contigs, with a similar observation in rye [2.1]. Analyses illustrate that 67% of discarded positions aggregate in contigs shorter than the L50 (1,425 bp) and 66% have a distance to the contig end that is below 200 bp. According to previous findings that stated terminal SNVs as less reliable compared to non-terminal SNVs [179], our decision was similar and we excluded them from further analysis. In total more than 15 million unique SNVs were identified, having a robust SNV quality score that exceeds a value of 50 (Table 2). The high divergence of cultivated barley lines was emphasized by the large number of SNVs identified in the wild barley *H. vulgare ssp. spontaneum* (6,191,130 SNVs in chromosome arm assigned contigs). The number of SNVs was more than twofold compared to cultivated barley accessions. In these accessions an average number of 2,742,630 SNVs in chromosome arm assigned contigs was observed, ranging from 948,722 ('Haruna Nijo') to 3,651,330 ('Barke'). The benefit of a well annotated genome reference sequence with gene model predictions was utilized to further filter for SNVs in coding regions. On global average per genotype 6.68% or in absolute numbers up to 370,000 of the detected SNVs are localized in exons [173]. Among them, 50% were integrated into the published genetic/physical framework of the barley genome. With this, a genome-wide SNV

resource was provided that is applicable for large-scale analysis (e.g. map-based cloning) and that assists new genome-assisted breeding strategies [180]. In conclusion, publication [2.4] provided in-depth analyses of the genome architecture together with the first detailed description of diversity patterns in the barley genome.

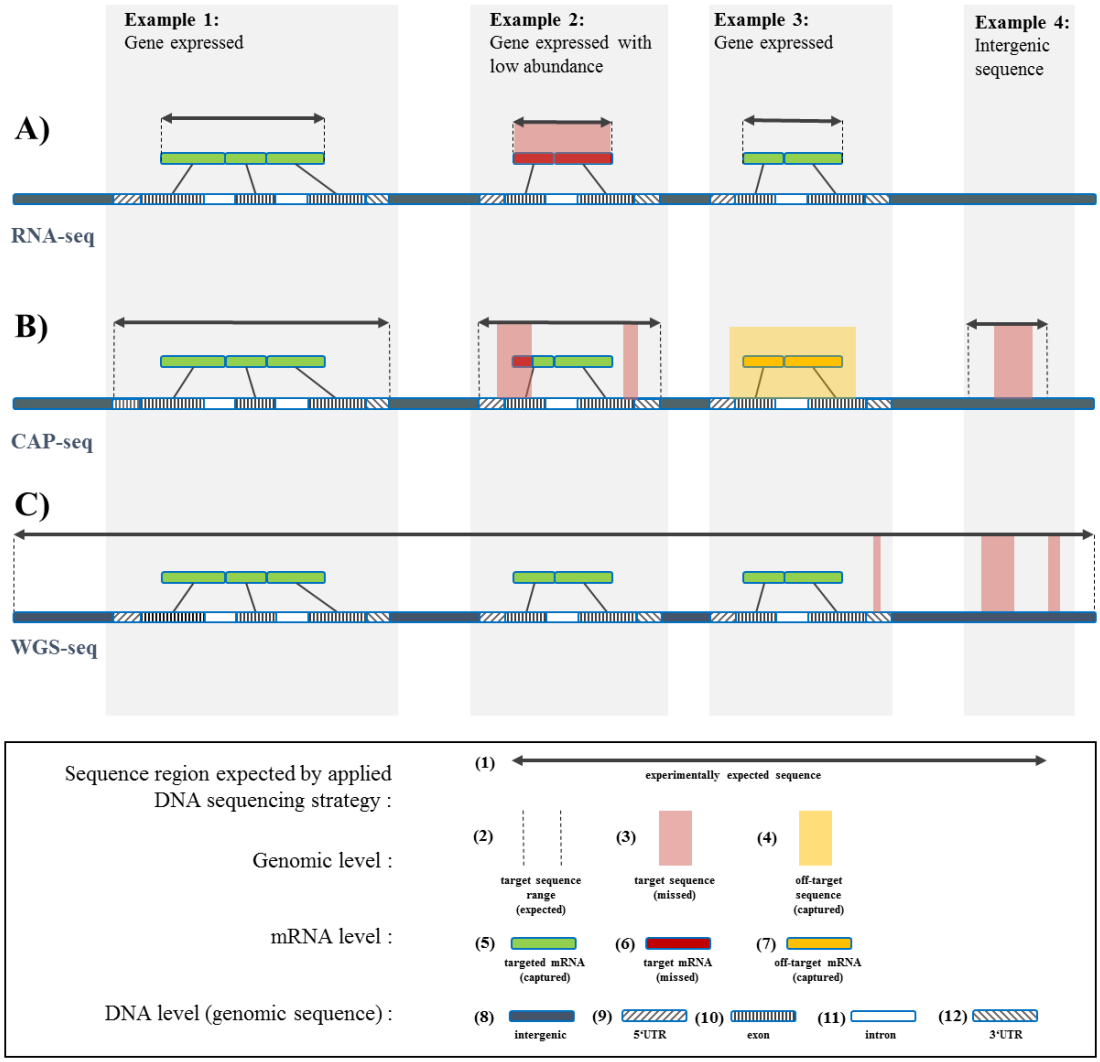

Figure 6 Comparative schema of predominantly accessed target regions. The Illustration shows with four simplified examples which genome segments are derived by each individual DNA sequencing strategy and what similarities and differences can be observed. For each of the three DNA sequencing methods A) RNA-seq, B) CAP-seq and C) WGS-seq the examples reflect the following simplified situations: 1) a gene with sufficient expression, 2) a gene having a low abundance transcript, 3) a gene with sufficient expression (off-target) and 4) an intergenic sequence. All depicted DNA sequencing strategies have a top segment (1) that shows the expected sequence and a bottom segment that is an abstract representation of the genome sequence. It has intergenic, exon, intron and UTR segments (8-12). Sections demarcated with (2) illustrate those genomic sequences that are expected by the particular DNA sequencing strategy. Additional sub-sections marked with (3) represent sequences that are expected as target sequence but were missed or (4) that were not expected as target sequences but were captured additionally (off-target). In addition, the mRNA level is illustrated with (5-7). The missed regions (3 and 6) illustrate the discrepancy between theoretical expected and experimentally achieved sequences.

4.1.2. Comparison of DNA sequencing strategies

In this section I will illustrate how each of the projects included in this thesis [2.1, 2.2 and 2.4] applied different DNA sequencing strategies and discuss advantages and disadvantages.

Comparability and predominantly assessed regions

The three DNA sequencing strategies clearly differentiate in the regions that are addressed and that, in consequence, are utilized for diversity analysis. RNA-seq is restricted to the coding regions of genes and, in comparison to CAP-seq, these are required to be expressed in the dissected tissue of interest. An advantage of CAP-seq is that through the customizable design, complete gene sequences can be targeted. Instead of assessing only the exons, introns and UTR regions can be included [181]. These differences are indicated in Figure 6, illustrating which part of the gene or genome is derived by each sequencing concept.
Discrepancies can occur between theoretical expected sequences and the experimentally achieved results. In RNA-seq, genes can be missed that are not expressed or only with low abundance [65]. Furthermore, this GCR strategy does only access the transcriptome (Figure 6, example 4 not captured). In CAP-seq, custom oligonucleotides can be used to access the full genomic sequence. However, as shown in example 2 and 4 sequences can be missed. This can be caused, when targeted sequences do not hybridize to the designed oligonucleotides [76]. Furthermore, parts can be missed because no oligonucleotides could be designed for the particular sequence (e.g. very low probe specificity). Another effect is illustrated in example 3, where sequences are captured that are not expected [182]. Here, sequences with high similarity to the designed probes are derived additionally (e.g. paralogous genes). WGS sequencing is not affected by these limitations (gene expression, oligo design problems or cross-hybridization) because no genome complexity reduction is performed. The complete DNA sequence of the genome is accessed. However, there are limitations like GC-bias, where GC-rich and AT-rich DNA fragments are underrepresented in sequencing results [48,183], that can lead to missing parts of the genome sequence (example 4). The discovery of diversity is only possible in regions that are derived by the applied DNA sequencing strategy.

Furthermore, it is important that the predominantly accessed sequences are comparable between different genotype samples. Due to various effects, like differences of gene expression levels in RNA-seq experiments or different sequence similarities of genotypes to the designed sequence capture oligos in CAP-seq, the coverage of read data in consequence is not equally distributed. Although the concept of CAP-seq is to target selected regions, the results of our analysis showed that

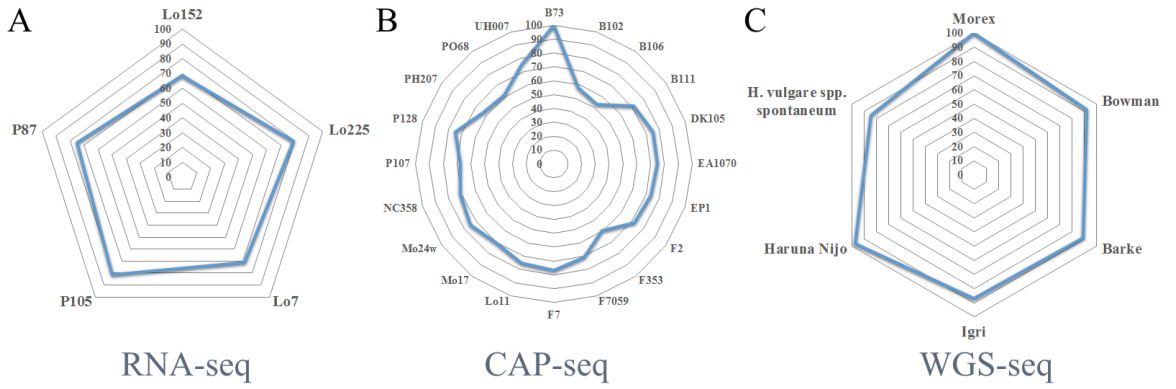

Figure 7 Comparability of coverage among different DNA sequencing strategies. For three investigated DNA sequencing concepts the percentage of reference genome sequenced is shown per individual genotype. The term reference genome in RNA-seq refers to the transcriptome reference and in CAP-seq it refers to the captured target-genes **A) RNA-seq in rye.** Different expression levels per genotype lead to variability. **B) CAP-seq in maize.** The effectiveness of capture oligo sequences differs between various maize inbred lines**. C) WGS-seq in barley.** Moderate variability with high average percentage of reference genome sequenced per genotype.

beside the majority of SNVs that locate on target regions (71%), also a substantial proportion of SNVs were detected off-target (29%). This indicates, that in CAP-seq experiments diversity is detectable in off-target regions, if these regions have sufficiently high sequence similarity to the designed oligo sequences. The sequencing coverage thereby is influenced by the hybridization reaction of capture-oligo sequences [76]. Analyses revealed [2.2] that even in custom designed capture assays the resulting output has high variability and is seriously affected by the studied genotype. For genotypes that are more diverse to the B73 reference line, which was used in [2.2] to design the capture-array, the hybridization to probes and thus the effectiveness of captured target DNA is lower. The experiments in maize generated significant differences in the total sequencing output, ranging at minimum from 149 Mbp (PH207) to 399 Mbp (B73). As expected, the reference genotype B73 yielded the highest efficiency (Figure 7B). The analysis investigated for each studied genotype, which percentage of reference sequence is sufficiently covered. The resulting proportions showed different representation per genotype (Figure 7), where 100% refers to the reference regions addressed by each of the three genome complexity reduction methods (transcriptome, target genes and genome). The transcriptome reference sequence for rye was constructed as a *de novo* RNA-seq assembly using five different genotypes. This might explain, why none of the five rye lines shows complete coverage of the reference sequence. In RNA-seq (Figure 7A), as well as CAP-seq (Figure 7B), the observed differences of sequencing coverage levels are partially restricting the comparability of diversity results among different samples. This can be caused when genotypes do not express particular genes (RNA-seq). Another reason is the hybridization effectiveness of the designed oligo sequences, when the capture array is applied to different genotypes (CAP-seq). This limits the

capture and in consequence, missing values dilute the global analysis. This effect is lower in diversity studies using WGS sequence data (Figure 7C), because no bias is introduced by expression or hybridization limitations. Interestingly, the WGS sequencing for cultivar 'Haruna Nijo', that was derived by single end 454 sequencing on lower sequence depth (~5-fold), resulted in a very high coverage of the reference genome (96.96%). The other barley cultivars, including the wild barley *H. vulgare ssp. spontaneum*, were sequenced with Illumina paired end sequencing with higher coverage, also resulting in sufficient coverage of the reference. However, the advantage of longer read length, at least partially, compensates for the higher sequencing cost of the 454 platform. The 'Haruna Nijo' WGS sample emphasized, that 454 sequences have a very high effectiveness and thus can be successfully used in diversity studies, too.
For a final comparison, the average coverage among the studied genotypes was calculated for each of the three evaluated DNA sequencing strategies. The proportion of the reference sequence (target region) which is covered by reads was evaluated for different samples (genotypes/cultivars). The results of this comparison revealed that WGS sequencing is better than RNA-seq or CAP-seq, achieving a high average sequencing coverage of 90%, 75% and 71%, respectively (Figure 7).

One of the major benefits of WGS is that it captures the complete diversity of a species. The resulting high density of SNVs cannot be recognized by RNA-seq, nor by CAP-seq. In consequence, with the established genome-wide SNV resource for barley [2.4], a new SNV density estimation of ~1/400 bp was proposed (referring to 4.6 million SNVs of cultivar 'Bowman' and an assembled reference genome size of 1.9 Gbp). This allowed us to correct a previous estimation of 1 SNV per 78bp [184], which was based on RNA-seq data of 23 genes. However, as shown in publication [2.4], the nature of such a WGS assembly reference is still fragmented. In consequence, WGS assemblies bear certain limitations (e.g. assembly errors) compared to a fully established genome reference sequence. This requires a careful investigation and filtering of discovered SNVs.

Criteria to assess the quality of variant positions

For a detailed assessment of the quality and reliability of predicted VPs an eligible set of criteria is required to filter the raw diversity data. The crucial ambition is to distinguish true variant sites from inaccurately called sites. In data sets without an established complete genome reference sequence (rye and barley), it is of great importance to evaluate the sequence environment of SNVs. For instance it can be analyzed if VPs locate in collapsed contigs, repeats or distance to contig ends. In [2.1], a manual inspection of discovered variant positions revealed a large proportion of erroneous variant sites towards the end of contigs. This finding also was observed in the diversity study of barley, using

WGS sequence data [2.4]. Low confidence variant positions were predominantly detected in small WGS contigs and in addition these, erroneous calls were found towards the end of contigs. This phenomena is widely observed in *de novo* genome assemblies of repetitive species [178]. In barley, it was shown for cultivar 'Bowman', that in contrast to the low-confidence sites, the high-confidence sites had an almost 4-fold higher occurrence in contigs, which were longer than the L50 contig length (~1.4 kbp). And in line with this, a 2.5-fold lower proportion of VPs with high confidence was observed in close distance to contig ends. Thus, to reveal a sufficient quality of diversity data, small contig and especially the corresponding contig ends need a careful investigation.

Analysis of read coverage at VPs is an expedient method to remove a significant amount of erroneous variant calls. This method is applicable to all three DNA sequencing strategies. However, it should be noted that the read coverage particularly in RNA-seq data is not equally distributed due to different expression levels (in case of un-normalized RNA-seq data). Several fundamental criteria can be utilized for basic filtering, using tools like vcfutils.pl, which belongs to the SAMtools package [153,154], or VCFtools [144]. In addition, several other criteria exist, e.g. distance to neighboring SNVs, minor allele frequency (MAF), distance to contig ends or density of SNVs at variant position. I have implemented the custom script VCF_filter.pl, which frequently is extended by further criteria and assists to determine robust VPs in diversity studies. It was applied the first time in the diversity study of the apomictic plant *Ranunculus* [185]. These and other developed methods were consistently improved and their applied results were integrated in several publications [185,186]. With this, I provided a continuous and sustainable contribution to the plant genomic and diversity research community.

Functional annotation of variant positions

The functional annotation of discovered diversity data provides additional information by the classification of VPs in different subclasses e.g. synonymous or non-synonymous VPs. The analysis requires the knowledge of gene models from the reference sequence. In maize [3] and barley [173] these required resources were established in 2009 and 2012, respectively. In species like rye, that are lacking a curated reference sequence or that only have a RNA-seq assembled transcriptome reference, this annotation process is not directly applicable. An indirect alternative could be to use gene information of closely related species [187]. In that respect, the WGS resources established in barley and maize are beneficial for a conclusive annotation. On the other hand, the diversity discovery in the rye project [2.1] relies purely on RNA-seq data. The RNA-seq concept intrinsically pinpoints to variants in exon sequences and thus, a substantial proportion of detected VPs is expected

to have an effect at the protein level [166]. However, it is possible to predict open reading frames (ORF) with tools like 'ORF Finder' (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi), OrfPredictor [189] or *get_orf* from the EMBOS suite [190], but their estimation of the correct reading frame is less reliable. In consequence, we did not apply a functional annotation of VPs in the rye transcriptome data.

An example that shows which additional information can be achieved by the functional annotation of VPs is given in the maize study [2.2]. Here, I analyzed the discovered diversity data in more detail using the gene model information of the reference genome. By this, direct gene effects were observed. The majority of SNVs (78.1%) were silent mutations, enclosed in intergenic or intronic sequences. However, a substantial number of SNVs (42,685) were identified as non-synonymous mutations, with 8,228 SNVs classified as non-conservative missense and 34,457 SNVs as conservative missense. Understanding the effects of discovered diversity sites by a functional annotation can potentially provide important arguments that can support research hypotheses. This was outlined with the publication [2.2]. Here, the information of a well-established SNV annotation identified a condensed number of candidate genes (235) that were affected by non-synonymous variants and these candidates can be used for further analysis in biomass-related studies.

Discovery of informative SNVs

The final set of discovered variants comprised 17,917 VPs in rye, 383,145 VPs in maize, and over 15 million VPs in barley (Table 2). Subsequent characterizations of VPs are needed to reveal which SNVs are informative throughout a population or germplasm collection. It is especially advantageous in the design of a genotyping array to know if a SNV will be informative throughout a larger panel of genotyping. The polymorphic information content (PIC) is a measurement of heterozygosity and initially was defined by Botstein et al. [119]. It is an indicative measurement to prevent ascertainment biases in the design of genotyping assays. The ascertainment bias is a systematic distortion in measuring the true frequency of a phenomenon (e.g. SNP) due to the way the data collection or sampling process was performed [191]. As follows, one should avoid selecting SNP markers which solely occur in elite lines and have very low PIC values. As a result, the selected high quality SNVs have an increased probability of successful divergence when used for genotyping. When multiple genotypes are in concordance for a predicted variant allele, the probability increases, that this particular reference position is a true positive SNV. Consequently, the reliability of predicted VPs increases with the number of genotypes included in a study. Therefore, DNA sequencing strategies, that use a genome complexity reduction (RNA-seq and CAP-seq) and thus can be adapted more easily to a large number of genotypes, have an advantage. VPs supported by multiple genotypes can

be prioritized.

In the maize project [2.2], we identified almost 13,000 SNVs in exons and these were classified as high confidence, because their detection was independently verified in more than 5 inbred lines. However, the majority (86.3%) of coding SNVs (86,875) were classified as rare variants, predicted in 5 or fewer inbred lines. This confirmed that maize is a species with an extremely high variability, as revealed in previous studies [192]. A large proportion of these rare variant positions was validated as true positive SNVs [193], overlapping published maize diversity sets e.g. the maize HapMap version2 [194]. In conclusion, a large sample size contributes to the global reliability of the variant detection.

Alternative strategies applicable for diversity studies

The diversity of large genotype collections can be assessed by the application of existing genotyping assays instead of performing an *in silico* diversity detection. In barley, a collection of 2,417 accessions from the USDA National Small Grains Collection was assessed by Muñoz-Amatriaín et al. [195] utilizing an existing genotyping assay. It comprises 7,842 SNP markers from the iSelect platform. A similar assay is available for genotyping in maize, including 49,585 SNP markers [128]. In conclusion, genotyping assays provide a cost-effective means to access plant genomes because these chips are applicable for large-scale analysis. However, the initial design is cost-intensive and the assays represent only a selective subset of the full diversity of a species. In the maize project [2.2] for instance, not all of the 4,648 selected target genes are represented by a SNP marker on the genotyping chip. Furthermore, the selection often represent SNP markers for certain traits in particular elite lines and this results in ascertainment bias [196]. Therefore, targeted sequencing of individual genomes is expected to alleviate this bias and detect rare functional variants, in contrast to genotyping assays. A further alternative for genotyping is the GBS approach, which combines sequencing and genotyping, without the requirement of designing a genotyping chip [125]. With the successful application of GBS in many plant species, e.g. maize [125] and barley [197], the method offers promising prospects.

4.2. Evaluation of computational methods

For a reliable discovery of variant sites, a high level of accuracy is required in all sub tasks and computational methods. This section will give a brief overview of the evaluation of different tools for read alignment and subsequent variant calling. The majority of these results were described in the publication [2.2].

4.2.1. Impact of read alignment

A high-quality read alignment is a crucial prerequisite for variant discovery. Thus, read alignment is of importance because its results directly influence all downstream analyses (e.g. variant calling). In comparison to Illumina derived sequence reads, that have been included in various evaluations of read alignment tools [198], the performance of 454 sequences in a subsequent read alignment is less carefully studied. This analysis on 454 sequence data was pending and motivated the performed evaluation. In that respect, I evaluated seven read alignment tools ('mapper') to choose the optimal alignment for variant discovery. The analyses were reported in publication [2.2], using 454 sequence data of 21 maize inbred lines. The comparison used important performance parameters including proportion of mapped reads, percentage of reference position with aligned reads, and the percentage of reads aligned on-target. The best alignment was produced by the tool BWA mem [199]. The tool Stampy [200] constructed a read alignment, which was close to the best performance of BWA mem.

In addition, the impact of read alignment within diversity studies was estimated. I performed and analyzed a total of 504 combination of read alignment and variant calling results, published in [2.2]. In this detailed investigation, for each of the eight variant calling methods, the number of successfully detected true positive VPs was compared to results of the same variant caller, using a different read alignment method. This allowed us to estimate the impact of read alignment on subsequent variant calling. The observed variability (detected range between minimum and maximum number of true positive sites for a particular variant caller) was, on average, 16.41% (maximum 26.15%) per variant calling method. This led to the conclusion that applying a thriving read alignment method is of high importance and methods with lower performance might cause a substantial loss of VPs.

Confidence of variant predication

Different methods have been applied to gain confidence of predicted variant positions. In many cases, wet-lab methods are required to validate and assess the sensitivity of variant prediction. However, this type of validation requires manual inspection, which is not practical for the entire set of detected variants, especially when considering large datasets. Secondly, a validation is possible by comparing to existing genotyping assays. Although this is a common practice in many species e.g. maize [201], only available assays are feasible, because a custom design of genotyping assay is expensive [148]. However, the tremendous advance of genotyping technologies towards large scale has made several of these validation sets available. Comparing *de novo* detected VPs against validated polymorphic sites allows for the use of measurements such as sensitivity, specificity, and F1-score, as applied in

[2.2]. In brief, sensitivity assesses the power of a tool to detect true positive sites. On the other hand, specificity estimates the true negative detection rate, measuring the ability of a method to discard true negative calls. The F1-score is the harmonic mean, balancing precision and sensitivity. In barley [2.4], detected variants were validated by two independent methods. The first was done by applying capillary re-sequencing to a subset of 300 SNV positions. 230 were assessable by the re-sequencing, and for 220 (96%) the validation succeeded. Second validation was achieved by utilizing a set of 3,972 verified barley iSelect SNP-markers [202]. The concordance between the assay and the detected VPs was determined by aligning SNP marker sequences to the WGS assembly contigs of barley reference cultivar Morex. In 87% of the anchored SNP markers we found consistency, assigning a high sensitivity for the applied variant calling in the barley genome.

In maize, being one of the most economically important crop plants, four large diversity data sets are available to assess confidence of the variant calling [2.2]. These resources are comprised of: 1) the 50k SNP genotyping array [128], 2) a GBS-derived control [125], a RNA-seq based discovery of 3.1 million markers [201], and the largest set 4) the maize HapMap2 [194]. In rye, the publication [2.1] designed the first large scale genotyping array (Rye5k) and thus provided itself a valuable resource for verification. However, in preparation for the design of the array a manual inspection ensured further confidence of selected variants. The final assay was successfully applied for 59 inbred lines, providing valid signals in 60% of the 5,234 SNP markers.

4.2.2. Comparative analysis of variant calling methods

Several different methods for diversity calling have been utilized throughout the publications of this thesis. Their broad evaluation of performance is the logical consequence. Therefore, publication [2.2] incorporated an extended list of variant calling programs, whereas Freebayes (previously named PolyBayes) and SAMtools represent methods that were applied in the other embedded publications [2.1 & 2.4]. Although each research project requests individual adaption, a schematic workflow exists to process NGS sequence data and reveal the genetic diversity. The workflow can be divided into four major processing blocks: 1) Base quality assessment, 2) read alignment, 3) variant discovery, and 4) variant filtering and annotation. One has to be aware that many downstream analysis tools are not capable of adapting to low-quality sequences. Therefore, quality trimming should be applied prior to computational analysis to avoid misinformed biological conclusions. Another important strategy to improve reliability of subsequent variant calling is the removal of redundant reads that are technical duplicates [203]. DNA sequencing machines produce substantial amounts of these duplicates [43]. Thus, removal of duplicates is required to avoid skewed downstream analysis and to decrease the likelihood of false positives in variant calling [204].

Table 3 Additional information provided in extended VCF format.

Field	Name	Type	Description
VC	Variant caller	String	List of tools
VCC	Variant caller count	Integer	Number of tools that detected particular variant site
VCD	Variant caller depth	Integer	Field is a list providing for each of the tools the observed read coverage (consistent order to VC field)
MD	Maximal depth	Integer	Maximal detected read depth
VCRD	Variant caller reference depth	Integer	Field is a list providing for each tool the number of reads with reference allele (consistent order to VC field)
VCAD	Variant caller alternative depth	Integer	Field is a list providing for each tool the number of reads supporting non-reference (alternative) allele (consistent order to VC field)

For the purpose of a detailed and comparative investigation, I implemented a framework system to automate the process of read alignment and variant discovery. In addition, parsing methods were developed to integrate and combine all generated results into a consistent and homogenous data structure that relies on the VCF format. With this approach, a tool-independent and uniform comparison is ensured. Finally, a VCF file with six additional data fields (VC, VCC, VCD, MD, VCRD and VCAD) was created that assess the joint information of all variant calling tools (Table 3). The developed concept was successfully applied in publication [2.2]. As emphasized by Wu and Cui [205], efficient and accurate SNV calling is required to avoid diluting effects of erroneous calls and to conclude solid hypothesis from achieved results. Following this recommendation, multiple variant calling results were integrated into one combined data set and this was used to compare the independent predictions and to assess their confidence.

The eight applied variant calling methods were further analyzed to evaluate their performance, taking into account important accuracy measurements. Therefore, I used verified SNV positions from published diversity data of maize (50k, GBS, RNA-seq and HapMap2) and extracted the overlap to our maize target regions [2.2]. Using these validated sets of variant positions, I calculated sensitivity, specificity, and F_1 score for the eight variant calling methods (SAMtools, VarScan2, CRISP, CLC find_variations, FaSD, SNVer, VCMM, and Freebayes), the highest average sensitivity among all the external control datasets was observed in SAMtools (0.94). Highest specificity was achieved by SNVer (0.52), and the best F_1 score was achieved by the tool FaSD (0.42). Furthermore, for each variant calling method it was calculated, which proportion of the final set of VPs (383,145) was detected. This final check revealed that SAMtools performed best, identifying 93.8% of the high-confident VPs. The good performance of SAMtools was shown in several other publication [150,206].

However, these results emphasized a considerable variability in the performance of different variant

caller. These finding are in line with the described low concordance rates observed in previous publications [146,155]. A method to utilize these differences for an improvement is developed and will be presented in the upcoming section of this thesis [4.3.2].

4.3. Improvement of variant calling accuracy

Construction of a genome reference sequence became feasible with the advent of advanced sequencing technologies even for complex plant genomes. In 2012, the IBSC published the first genetically and physically anchored sequence of the barley genome [173], which guided research onto new levels. Nevertheless, WGS assemblies also established by many plant consortia often do not reach high quality standards in terms of full representation of a genome [178]. Fragmentation, lack of anchoring, and ordering of contigs together with small and collapsed repeat contigs are inherent characteristics of these reference sequences. Variant calling in these references is widely used can provide novel insights. Nevertheless, WGS assemblies need careful consideration when diversity studies are performed, since they can be the source for erroneously called SNVs [207]. The publication [2.3] describes the tool Kmasker, which aims to detect single or low-copy regions in the highly repetitive genome of barley. Detection and removal of repetitive regions is crucial for variant calling, however, it is usually not feasible to remove these regions with a common filter based on low-coverage. This emphasizes the need to establish a method that does not rely on read alignment. Subsequent removal of putative false positives will lead to an improvement of accuracy, as shown in the following section [4.3.1]. In addition, the aim of decreasing the rate of erroneous variant calls is further supported with the proposed concept of combinatorial variant calling (CVC), introduced in section [4.3.2] and published in [2.2].

4.3.1. Enhanced accuracy through repeat investigation

Repetitive sequences permeate the genomes of species at all levels of the tree of life [170]. High percentages of repeats are detected in genomes of crop plants [208,209], with over 80% in maize [3], 84% in barley [173], and approximately 92% in rye [210]. Reliable detection of these repetitive sequences has been emphasized as a requirement to avoid false positive variant predictions.
In recent publications, repeats, also referred as low-complexity regions, are identified as one of the major causes of erroneous variant calls when using WGS reference sequence [211]. In crop plant research, a feasible method to overcome this challenge is to use manually curated repeat databases (e.g. the Triticeae repeat composition TREP) [212] to mask known repeats based on sequence similarity. Programs like RepeatMasker [213] or DustMasker [214] can be utilized for this purpose. A major disadvantage of this approach is the need for a library that includes known repeat elements

a requirement often not met for many species. Automated construction of repeat libraries can overcome this, for instance by utilizing the tool RepARK [215]. With the tool Kmasker [2.3], I implemented another alternative that counteracts the uncertainty of repeat-derived false positive SNVs. It masks repetitive sequences and discovers the single- and low-copy regions within barley sequences. This concept relies on *k*-mer counting and is applicable to any species. Utilizing *k*-mer methods is beneficial, because even the *de novo* detection of yet unknown repeats is feasible. Especially in species with incomplete genome sequences like rye and barley, this feature is of decisive importance. Additionally, *k*-mer methods are most relevant in WGS sequencing, where the assemblies often contain collapsed repeat regions [170]. That is, repeat sequences from different origins in the genome aggregate into single contigs representing the repeat motif [170]. In a subsequent read alignment, repetitive reads are assigned to their best alignment positions, leading to a clustering of reads from different genomic origin. These collapsed repeat regions are the source of many false positive SNVs. Therefore, to overcome this error it is necessary to identify these repetitive sequences before variant calling.

The utility of Kmasker was shown on large scale by using the discovered diversity resource of barley [2.4]. Here, the discovered set of putative SNVs comprised a large amount of ambiguous positions that were removed (Table 2). First, these discarded positions were analyzed in more detail for barley cultivar Bowman (including 50 bp upstream and downstream sequence of a VP).

In Bowman, more than 12 million positions were removed as ambiguous (e.g. heterozygous) positions. Small contig sizes and a high proportion of putative VPs towards the end of contigs are widely observed sequence characteristics of these rejected sites. In ~77.6% of the discarded contigs and 68.9% of all contigs, the average *k*-mer frequency at the contig ends is significantly increased (>5-fold), compared to the overall *k*-mer frequency of the corresponding contig. Furthermore, a considerable proportion of the discarded positions (~18.8%) are characterized by very high average *k*-mer frequencies (>50). The observed results indicate that the repetitive nature of the barley genome lead to the fragmented and collapsed assembly, also observed in WGS assemblies of other species with complex genomes [216]. The use of Kmasker provided an independent confirmation that the majority of the ambiguous variant positions should be discarded. In conclusion, additional confidence can be assigned to remaining VPs.

In a second analysis, I investigated the *k*-mer spectra of the discovered 'Bowman' VPs. The *k*-mer spectra is the distribution of all DNA *k*-mers of a species [217] and the analysis of the frequencies of these *k*-mers provide a useful perspective to the complexity of a genome, especially in large and repetitive species like barley. The comparisons involved raw and filtered VPs, as depicted in Table 4. For each of the six *k*-mer frequency classes, the proportions in each diversity set are given. The analysis of the *k*-mer spectra in coding sequences revealed that the majority of SNVs in coding

Table 4 *K*-mer spectra in barley cultivar Bowman (*k*=21). The detected variant positions were classified by four filtering strategies ('raw', 'basic', 'depth' and 'exon'). Subsequent *k*-mer analysis was performed for each data set to calculate the quantity and proportion (%) of *k*-mer class ranging from 'low' to 'extreme'.

k -mer class	*k* -mer frequency (range)	Bowman[1]	%	Bowman[2] (basic)	%	Bowman[3] (depth)	%	Bowman[4] (exon)	%
total		10,064,822		2,060,992		1,791,554		11,690	
ambigious	0	54,149	0.54	7,436	0.36	4,646	0.26	54	0.46
low	>0 to <10	6,254,421	62.14	1,171,050	56.82	1,034,306	57.73	11,544	98.75
normal	>=10 to <20	914,896	9.09	207,269	10.06	173,960	9.71	41	0.35
medium	>=20 to <50	1,127,180	11.20	248,812	12.07	207,763	11.60	14	0.12
high	>=50 to <200	1,158,235	11.51	264,297	12.82	225,942	12.61	33	0.28
extreme	>=200	555,941	5.52	162,128	7.87	144,937	8.09	4	0.03

[1] complete set of variants. Included VPs have sufficient distance to the contigs ends (>50bp) and the enclosing 100bp reference sequence does not have undefined ('N') nucleotides
[2] basic filtering. Rejection criteria: 'N', score <50 and ambiguous position in 'Morex' WGS assembly reference
[3] depth filtering. Rejection criteria: 'N', score <50, ambiguous position in 'Morex' WGS assembly reference and depth <10 reads
[4] exon filtering. Rejection criteria: 'N', score <50, ambiguous position in 'Morex' WGS assembly reference and position overlapping coding sequence of a gene (exon)

sequence are characterized by low complexity (*k*-mer frequency score <10). This correlation has been shown before [218]. However, the results show that basic filtering strategies do not have the ability to fully capture the repeat complexity. When comparing raw and filtered diversity sets (Table 4), the filtered sets still contain a large proportion of positions that originate from regions classified as high or even extremely repetitive. Basic filtering (e.g. read coverage) is important and revealed a solid and qualitative diversity resource [2.4]. However, the analysis of repetitive characteristics at VPs was not a conscious filtering criteria in the applied diversity pipeline. As a result, the ratios of high *k*-mer frequency did not change in the filtered data sets (‘basic’ and ‘depth’). These findings are in line with the recommendations of Kurtz et al. [50], that the identification of repeats with *k*-mer methods is beneficial for improved downstream analysis. However, all studied barley cultivars still have a high proportion of SNVs that are enclosed in sequences with medium or high repetitive *k*-mer sequences (~30%), which require additional attentiveness. This necessary caution can be supported by the tool Kmasker.

In conclusion, Kmasker assisted in distinguishing true variant calls from false positive variant calls. Furthermore, the tool was successfully applied in a TALEN approach in barley to evaluate potential off-target cleavage [219]. However, the investigation of *k*-mer counts at identified VPs presents a yet unexploited potential. Here, the sequences that encapsulate VPs could be analyzed for repetitive features or characteristics of mis-assemblies [220,221]. Kmasker could partially close this gap. Continuation of these studies would be very beneficial, for instance by evaluating SNP markers from large genotyping assays. Thereby, *k*-mer analysis could be used to compare successful versus less optimal performing SNP markers.

4.3.2. Combinatorial variant calling approach

The standard procedure to detect the genetic diversity of a species usually relies on a single variant calling method [222–224]. However, the application of multiple variant calling methods can eliminate a considerable amount of false positive variant sites by detaching skewed candidate positions. Recent comparisons of variant calling methods [151,155] agree with the findings of publication [2.2]: considerable discordance is observed within different variant calling tools. Therefore, I proposed and developed the concept of 'combinatorial variant calling', to establish more reliable predictions. The discrepancy between individual variant calling programs is because each one has a different implementation [151] and this leads to marginal or even clear differences [225]. In addition, results can deviate even further when different optimizations of internal thresholds are applied (e.g. base quality). Therefore, combining the predictions of multiple tools results in a sharp increase in the number of candidate positions. However, the combinatorial variant calling approach can decisively improve the accuracy. It uses the additional information of multiple tools to gain further confidence of a VP.

As shown in [2.2], utilization of multiple tools provided additional support and decreased the risk of erroneous calls. As a result, a significant reduction of candidate positions was observed. From an initial set of over 4.8 million detected candidates, a more robust number of 383,145 VPs was selected. These variants were predicted by multiple tools (at least three out of eight variant calling programs). To prove validity of the extended filtering, two additional tests were evaluated. First, the final identified 383,145 VPs were compared to the stand-alone application of each single variant caller using basic filtering. The basic filtering, together with the combination of multiple diversity sets from different variant callers, resulted in almost double the number of SNVs (681,993). This complete set was used as input for the CVC concept. To show the quality of the applied filtering the discarded SNVs (298,848) were analysed in more detail. The majority was observed by a single tool and only in one single genotype (56.0%). In addition, 24.1% were solely predicted by a single tool and another 11.2% were observed in a single genotype (11.2%) only. This shows the low confidence and emphasizes the need to discard these positions. To further confirm that these positions do not reflect true variants, a second test was performed. Therefore, the polymorphic information content (PIC) was calculated for the complete set of VPs. For the discarded positions, the calculation revealed a very low PIC value, with a median value of 0.07. A small fraction of 6,982 sites (2.33%) was characterized with sufficient PIC values (0.2-0.5). In contrast to that, from the finally selected VPs (383,145), more than 81,000 positions (21.6%) had a sufficiently high PIC value, demonstrating that these VPs are detected in multiple genotypes.

An alternative strategy to increase reliability of the variant calling, without using the CVC approach, is to utilize a more stringent filtering of read depth. To show the performance of this alternative strategy the required read depth at VPs was increased from 5 to 10, where 5 was the parameter setting used in [2.2]. A threshold of 10 reads per VP led to a clear reduction of detected VPs. In comparison to the final diversity set, detected by the CVC approach, only 31.6% of the VPs were detected with the alternative approach. Therefore, a clear disadvantage of the alternative method is to lose confident SNVs that are supported by multiple variant calling methods.

In summary, the presented evaluation verified the high quality of the CVC concept. It is beneficial in comparison to a stand-alone application of variant calling tools and results in improved reliability assigned to final diversity sets. It is a powerful option, especially when the diversity calling is performed in low-coverage sequence data. Beside the application in the maize projects [2.2], which is embedded in this thesis, I successfully applied the CVC approach on a large scale in *Brassica napus*. Utilizing the re-sequencing data of 52 lines, the diversity throughout the genome of *B. napus* was studied, revealing 4.3 million high quality SNVs [226]. With these results, that were revealed using various methods developed in this thesis, we established a novel and powerful diversity resource for *B. napus*.

Finally, it should be noted that the developed method neither involves an additional manual inspection, nor does it requires further laborious web-lab experiments. Furthermore, several other strategies can be considered to increase the accuracy in variant calling. A selected brief outline of additional improvement methods is given in the next section.

4.3.3. Alternative strategies

This section is a discussion of alternative computational methods that can be applied individually or in combination with those previously mentioned to improve the accuracy of diversity detection. Several alternative variant calling programs have had a broad application and have been used in multiple genome studies like GATK [168] and SOAPsnp [164]. However, SOAPsnp has the limitation that it only accepts read alignments constructed with SOAPalign and thus, was excluded from the comprehensive evaluation.

Furthermore, several other strategies can be applied prior the actual variant calling process. First, the use of error correction methods have recently gained widespread attention, due to the large amount of sequence data generated even for non-model species. High-throughput sequence data therefore is used to detect and correct possible sequence errors. Several tools were published in the last few years to address this problem with quite different performance results, as emphasized by Yang et al. [49] in a review of error correction methods. Standard correction strategies typically apply *k*-mer counting to differentiate between trusted and untrusted *k*-mers [227,228] or related methods using suffix trees

[229,230] or suffix arrays [231]. The common intention is to correct for errors at the nucleotide level by revealing inconsistencies in the *k*-mer spectra. Repeats increase the complexity in the correction process as emphasized by Yang et al. [232] and thus, successful application in highly repetitive crop plants was missing. A first evaluation was accomplished by Chris Ulpinnis in the framework of a bachelor thesis [233] that I co-supervised. The application on barley sequence data revealed only minor reduction of sequence errors. Furthermore, contrary effects were observed by the integration of additional sequence errors. These systematic errors throughout the correction process have serious effects on downstream analyses such as variant calling. In consequence, error correction in highly repetitive genomes, like maize, barley, and rye, must be considered carefully. However, it is expected that error correction will become increasingly important, especially in the context of long sequence read correction [227,234].

The second approach to increase the reliability of variant calling is by improving the read alignment accuracy. These methods are utilized after the actual read alignment process. Two general strategies are briefly introduced. For the first method of base quality calibration, several tools are published including GATK 're-calibrate' [235] and Novoalign [236]. Here, the associated base quality score is corrected after the mapping process. This intends to gain a better estimation of reliability by using the information provided by neighboring sequence reads. It is noteworthy that quality trimming methods initially applied to the raw sequence data have been shown to be advantageous [198]. At the end of these correction steps, the causality of low quality bases and putative false positive predictions might be, at least partially, reduced. The second strategy is to perform a local re-alignment of reads, also done after the read mapping. Here, the alignment information is investigated again, to search for a more optimal position of single or pairs of sequence reads. Tools like SRMA [237] or COVAL [238] aim to detect spurious alignments of reads and assist in a local re-alignment. In consequence, these adjusted read alignments can lead to an improved accuracy of variant calling. Enhancing the INDEL detection accuracy is also conducive for a diversity study. INDELs can have a serious impact, especially when located in the coding sequences of genes [239,240]. Therefore, high precision in the detection process is crucial. The integration of alternative programs is reasonable, because the introduced variant calling programs support the prediction of single nucleotide mutations or short INDELs, but lose prediction accuracy for longer INDELs. More accurate prediction can be achieved by applying methods that utilize the split read approach like SV-M [241]. Furthermore, an assembly based re-alignment of reads is applied by the tool ABRA [242] to improve the quality of INDEL detection.

Table 5 Methods for improvement and filtering of high quality variant sites. This overview is a continuation of the filter criteria given in Figure 5. Methods are listed with direct notification (3rd column) to the associated sections of this thesis. In addition, the 4th column directs to the associated publication. Some filter or methods were discussed in the frame of this thesis but were not applied throughout the four cumulative publications. Primary reason was a continuous development of the related filtering methods making certain methods applicable only for subsequent projects. All methods listed as 'IM' have to be applied prior the variant calling and thus are not directly linked to the actual process of diversity detection.

Filter	Type*	Associated Section	Associated Publication	Detailed Description
Homozygous or heterozygous	BF	4.3 4.4.1	2.1 2.2 2.4	Homozygous, if the alternative allele is present in <10% or >90% of reads, otherwise heterozygous
Read coverage	BF	1.5 4.1	2.1 2.2 2.4	Requirement of minimal read coverage to select for robust sequence position. Positions with low read coverage tend to originate from sequencing errors
Contigs length	BF	4.1	2.1 2.4	Removing contigs <200bp from WGS assemblies is a standard procedure to avoid variant calling in collapsed or irritated contigs
Distance to sequence end	BF	4.1	2.1 2.4	Sequence ends have higher risk to be biased by assembly problems and lack sufficient quality to be used for marker probe design
Quality score	BF	1.3 1.5	2.2 2.4	PHRED like score that is incorporating the mapping quality
Distance to homopolymer	BF	1.2.2	2.2	Homopolymer errors are mainly observed by 454 sequencing and SNVs detected close or overlapping these regions have higher risk to be false positive
Distance to ambiguous nucleotid	BF	4.1	unpubl. data	Ambiguous nucleotides e.g. 'N' or other non-reference positions not considered as valid SNV (e.g. sequencing errors)
Number of ambiguous nucleotids	BF	4.1	unpubl. data	Maximal number of ambiguous nucleotides in a 50bp / 100bp frame around a SNV; relevant for custom probe selection in marker design process
Neighborhood quality standard (NQS)	BF	1.5.1	2.1 2.2 2.4	NQS has been incorporated into majority of variant calling methods to calculate quality scores
Functional SNV annotation	EF	4.1 4.3.1	2.2	e.g. non-synonymous (nsSNV), synonymous (sSNV), splice site affecting SNVs, intergenic or intronic SNVs (silent mutation), premature stop or lost stop codon
K-mer repeat investigation	EF	4.1 4.3.1	2.3	*K*-mer frequency analysis to estimate risk of repeat derived false positive SNV calling
Combinatorial variant calling (CVC)	EF	4.2.2 4.3.2	2.2	Confidence of variant calling is increased by using multiple prediction methods. Including information of VC, VCC, VCD, MD, VCRD and VCAD (see Table 3)
Minor allele frequency (MAF)	LF	1.5.1 4.1 4.4.1	2.2	Appropriate filter criteria for large collections of genotypes to classify rare and widely spread alleles
Missing genotype rate	LF	4.1	2.2	Higher confidence for selected SNV data is achieved when low rates of missing genotypes (<3%) are applied
Polymorphic information content (PIC)	LF	4.1 4.3.2 4.4.1	2.2	Measurement estimating the usefulness of a marker, where high PIC values (1) correspond to an estimation of high marker performance and low (0) to a missing functionality
Base calibration	IM	4.3.2	outlook**	External tools e.g. GATK or Novoalign
Quality trimming	IM	1.5 4.2.2 4.3.2	2.1 2.2 2.4	External tools e.g. CLC quality trimming or FASTX-Toolkit
Duplicate read correction	IM	1.2.2 4.2.2	2.2 2.4	Duplicated reads can superimpose the correct allele frequency with possible consequence of erroneous variant calls.
Error correction	IM	4.3.2	outlook**	External tools e.g. Quake or HiTEC
Re-alignment	IM	4.3.2	outlook**	External tools e.g. SRMA or COVAL
INDEL correction	IM	4.3.2	outlook**	External tools e.g. SV-M or ABRA

* **BF basic filtering; EF extended filtering; LF large scale filtering; IM independent improvement method**

** **outlook of methods that have further potential to improve variant calling (pending evaluation).**

Access to the genetic diversity of a species requires accurate methods to reveal the true proportion of putative variant positions. In order to achieve the best precision in this demanding objective, it is necessary for each sub processes to achieve maximum accuracy. According to the recommendation of Guo et al. [243], the three main stages (raw data processing, read-alignment and variant calling) have to be conducted with proper quality. In this thesis I have applied, expanded, and developed various strategies, which increase the confidence of a variant prediction. The retrospective Table 5 depicts a conclusive overview for the most important filter, including corresponding sections of this thesis and related publications. The aim is to provide a summary and to direct the reader to relevant parts. Therefore, the three filtering classes of Figure 5 (BF, EF, and LF) are utilized again. The list is extended with the independent improvement methods (IM), discussed in this section. Diversity has many facets and equally diverse are the strategies for their accurate detection.

4.4. Scope of applications

The outlook of possible applications for discovered SNVs is increasingly entered by new developments, e.g. from the field of population genetics [244], or tremendous improvements within existing platforms (e.g. assay technology) that also are applicable for non-model species like crop plants with complex genomes [148]. In combination with the existence of reliable gene models for a studied species, the diversity resources can be used to determine, what, if any, functional effect VPs have on the gene level. This was shown in detail in publication [2.2]. In addition, VPs can be utilized for phylogenetic analysis to determine the relatedness within a population [2.2]. As emphasized in this thesis, DNA markers represent one of the most powerful approaches to analyze diversity in plant genomes at a broad scale. Through application of high-throughput marker technology, e.g. Illumina Infinium (http://www.illumina.com) or Affymetrix Axiom (http://www.affymetrix.com), heritable traits can be directly associated with its underlying genomic variation in the genome. However, interconnectivity in genomic analysis is present on multiple levels, as indicated by [52]. These close links of research fields imply that for instance diversity results do not only influence downstream genomic analysis, but also provide benefits for other scientific fields. For instance, by providing a decent control to improve the *de novo* assembled genome reference of a species. With the development of large genotyping assays, many research projects reached an important milestone for application in future research project. In the last decade, there was a paradigm shift from the use of SSR markers to SNP markers [245]. Because of its massive throughput, these efficient and cost-effective genetic tools became a breakthrough technology. Here, various genomic analysis and new breeding strategies are the prospective outlooks.

4.4.1. SNP marker development

One possible application of the discovered diversity is the development of SNP markers. The construction of large genotyping assays is an efficient approach, because it increases the density of genetic markers for an organism and consequently allows for the study of the genetic architecture of a trait or genomic locus in greater detail. In complex plant genomes, availability of genotyping assays provides vital information for the construction of linkage maps [246–248]. The construction of a high-density genetic map relies heavily on the high resolution of SNP markers within the genome. Because of their benefits, SNP markers are increasingly used in many species and the construction of a high-density genetic map represents an important milestone e.g. *B. napus* [249] or wheat [80]. The publication [2.1] successfully proved the direct conversion of discovered diversity into application by design and use of a genotyping panel (Rye5k). Emerging from this established diversity resource a genetic map was constructed. The determined genetic diversity evolved into a novel genomic toolbox for rye. Therefore, SNVs with the greatest likelihood of becoming a successful marker should be selected. The question is "What is the definition of an optimal marker?". Referring to Kumar et al. [250], the characteristics of an ideal DNA marker can be described by eight distinguishing features. In the following listing, each property is complemented with a brief explanation of how methods developed in this thesis support the qualitative selection of variants for marker design.

(1) Polymorphism – a successful SNP marker requires, that a detected VP is truly characterized by diverse alleles. In this thesis, I presented and developed strategies, to improve reliability of predicted VPs and to discard erroneous prediction.
(2) Stability – stable SNP markers are consistent and generate reproducible results. Markers should be stable among different genotypes and in repeated propagation of varieties. Hereby, the correct selection of VPs is a core objective (e.g. MAF is an important criteria). In [2.1], we showed successful portability of the developed Rye5k assay to related grass species like barley, wheat, and triticale.
(3) Co-dominance – is the ability to distinguish heterozygotes (call ratio 1:1) and homozygotes (call ratio 2:0) [251]. In general, co-dominant markers are more informative than dominant markers. In rye [2.1], maize [2.2], and barley [2.4] homozygous SNVs were selected.
(4) Cost efficiency – the massive reduction of costs for discovery of VPs and the availability of high-throughput genotyping assay led to an improvement in the cost of developing SNP markers. As described by Hiremath et al. [252], several different assay systems are available, offering investigators a flexible decision for

a particular system based on the number of SNVs and/or number of genotypes to analyze.

(5) Simplicity of discovery – cheap and reliable large-scale DNA sequencing facilitates their application for variant calling. Numerous variant calling methods support the discovery of SNVs and produce extensive amounts of possible SNP markers. SNP markers provide (simple) access to the genetic diversity of a species on broad scale. However, accuracy and reliability remain challenging objectives, as demonstrated in this thesis.

(6) Broad genome dispersion – high density of SNVs in the genome is advantageous, making SNVs a beneficial marker system. For maize [2.2] and barley [2.4], a distribution throughout all chromosomes was observed.

(7) Heritability – SNP markers can estimate heritability (e.g. of a trait) with a high degree of accuracy. Various estimators utilize SNP marker based methods to explain phenotypic variations [253,254].

(8) Reproducibility – the availability of various genotyping platforms that utilize detected SNVs on large-scale is beneficial. Robustness of genotyping results on different machines and different laboratories has been shown [255], assigning a very high reproducibility to SNP markers. In addition, high transferability between genotyping platforms has been shown [256].

The selection of reliable SNVs is crucial for the design of genotyping panels. In publication [2.1], I assembled the first transcriptome reference sequence and established a large-scale diversity resource for rye, which was ultimately converted into the Rye5k genotyping assay. In many plant species e.g. rice [138] or maize [128], such condensed panels of diversity clearly provide a benefit for scientists and breeding strategies. The application of these genotyping assays across species barriers provides an additional benefit, as shown in [2.1]. These cross-species applications can assign high reproducibility and are widely used as control for the developed genotyping assay [202,257] or more distant applications of capture assays [77]. In the context of reproducibility, the problem of ascertainment bias is an important issue. The development of genotyping assays that are based on a non-random samples or a less diverse selection sample (genotypes) of a population will very likely lead into an inadvertently skewed representation. In consequence, these marker sets will likely irritate genotyping analysis and skew phylogenetic relationships [258]. Resulting assays, including their design, often have a tendency towards elite lines and especially monitor domestication effects, but might lack to represent the full natural diversity. For example Frascaroli et al. [196] revealed a certain ascertainment bias in the Illumina MaizeSNP50 assay towards the North American dent germplasm. To avoid this bias in the design of genotyping assays, it is recommended to use large and diverse

genotype collections and in addition integrate a wild species or progenitor line to counteract the ascertainment bias.
In conclusion, SNPs have become the ideal marker system for many crops. The established diversity resources for rye, maize, and barley will further support a better understanding of the complex genomes of these crops.

4.4.2. Advantages for accelerated crop breeding strategies

In this section, an outline is given, which impact the established diversity resources have for crop breeding. Genome-wide association studies (GWAS) are a powerful tool to dissect quantitative traits and to pinpoint their genomic origin(s) [259]. It was successfully implemented in various economical important crops like maize [260,261] and barley [262]. However, for a precise association it requires a high density of DNA markers. Particularly in barley, the substantial increase of genome-wide SNVs, determined in this thesis [2.4], provides a highly beneficial advance for future studies.

In the context of crop breeding the genome-based breeding strategies supplement traditional breeding methods, which will accelerate crop improvement and allow for more precise breeding initiatives [263]. The availability of genome-wide SNP markers has significant impact for future crop improvements [264–266]. Years of cultivation and artificial selection have inevitably lead to a reduction of biodiversity in elite germplasm [267]. Therefore, large and reliable diversity resources are required that can be used to introduce new diversity into existing breeding programs [226,268]. To reach these objectives, genome-based strategies of marker assisted selection (MAS) and genomic selection (GS) are powerful tools [269,270]. On the one hand, their application will assist to facilitate new plant genetic resources and thus create new diversity resources. On the other hand, it is expected that the precision in the genomic selection process will result in testing fewer candidates in the field, because this is one factor to decrease cost and at the same time a major benefit of GS [271]. Hence, this selection for remarkable candidates, which are selected for the breeding programs only, can lead to a decisive reduction of genetic diversity. However, as depicted in rye [265] DNA markers are beneficial tools to systematically assess breeding pools and to comprise strategies for an advanced development of broad and promising germplasm. The established diversity resources [2.1, 2.2 and 2.4] and the improvements, in terms of accuracy and reliability of the discovery of variants, which I developed in this thesis, will have a continued impact for these new strategies of crop breeding.

Referring to the FAO, the world population will increase to an estimated number of 9 billion in 2050, requiring a 70% increase of food production [112]. Plant science is playing an essential role to

overcome this future challenge in nutrition supply [113,142,272]. Crop improvements will inevitably be linked to the application of modern breeding technologies [180,273]. Therefore, genetic diversity is the essential pillar and this thesis aims to gain robustness in their discovery and to provide new diversity resources.

5. Conclusion and outlook

High-throughput DNA sequencing technologies provide unprecedented opportunities to access the genome of a species. The high amounts of raw data production require a large number of computational calculations to translate pure read information into high-quality diversity information. This thesis illustrates how this can be achieved using different DNA sequencing strategies. These strategies represent the genome sequence of a species with different complexity reduction methods, ranging from the largest reduction of DNA sequence complexity (RNA-seq), to medium reduction (CAP-seq), or no reduction (WGS-seq). To reach the requirement of high accuracy in variant calling, each individual calculation has to be performed with the highest precision. With the combinatorial variant calling, I conducted a novel approach to improve *in silico* variant predictions at little or no additional costs. The second method of *k*-mer repeat investigation assists to decrease error-prone variant positions that are characterized by skewed *k*-mer frequency patterns. Both methods are applicable to all investigated DNA sequencing strategies and consequently, assist to improve accuracy of a diversity study. The comparative analysis of read alignment and variant calling methods revealed considerable differences, but also determined the best combined performance for the read alignment tool BWA mem and the variant calling tool SAMtools/VCFtools. However, in comparison to the application of a single variant calling method, the combinatorial variant calling approach improved the reliability of a variant detection, because the concordance of multiple tools increased the confidence of a prediction. Applicability has been proven in economically important crop plants. The embedded publications of this thesis contain three diversity resource, in whose construction I was considerably involved. A comprehensive investigation of the rye transcriptome provided novel diversity results. The study compiled into the publication of the first transcriptome reference sequence for rye, including a large set of variant positions, that were converted into a genotyping assay. For maize, a comprehensive set of gene candidates related to yield and biomass were investigated providing a new diversity resource. For barley, a genome-wide set of variant positions was delivered, providing novel insights of diversity. These established diversity resources will assist genome-based breeding strategies, such as marker assisted selection or genomic selection, leading to accelerated crop improvement.

Future challenges in genomics are hard to predict. Nevertheless high-throughput sequencing will be an integral part of many research projects and will shape more and more the aspects and goals of future science [171]. At present, sequencing vendors prefer to refine and improve their technologies

rather than release the next milestone breakthrough innovation [274]. However, with the continuous upscaling of throughput, DNA sequencing projects will increasingly reach the level of 'ultra-high' throughput sequencing with further tremendous increase of sequencing outcome. The bioinformatics community faces the challenge of improving computing speeds and data storage to keep pace with the DNA sequencing technologies. As a direct consequence, accuracy and quality checks are required to be performed in automated procedures because manual inspection or experimental validation will be infeasible at that throughput.

The steady development of new tools requires a standardized evaluation of current state-of-the-art methods for variant calling programs on a regular basis. To overcome this necessity, a proposed solution could be a 'Diversathon' initiative. The term is related to the 'Assemblathon' initiative [275,276], that aims to evaluate existing assembly programs on defined data sets. A similar concept could be conducted for variant calling methods. The conversion of the concept would be beneficial for the broader scientific community, because sequence and diversity data for more and more species will become available. However, the conducted procedures to determine variants often lack comparability. The massive increase of sequence data for virtually any species request semi- or fully-automatized analytical pipelines. Erratic automation in combination with a continuous overload bears the risk that quality of results is diluted. Standardized and reliable data sets would be required to assess the quality of diversity pipelines. A joint 'Diversathon' initiative is a perspective in which future diversity resources, particularly when constructed by highly automatized pipelines, can be compared to high quality gold-standards to achieve confidence and reliability.

6. Summary

Next-generation sequencing (NGS) is considered to be a breakthrough technology. In the course of its success, it can provide access to the genomic sequence of even large and complex plant genomes. Three major strategies exist to assess the genomic information of a species at different scales and complexity levels: transcriptome (RNA-seq), target capture (CAP-seq) and whole-genome shotgun (WGS-seq) sequencing. A comparison of their applicability in complex plant genomes was pending. The scope of this thesis was to evaluate each concept, ascertain its potential for determination of the genetic diversity, and develop methods for their improvement.

With these objectives, the economically important crops rye (*Secale cereale* L.), maize (*Zea mays* L.) and barley (*Hordeum vulgare* L.) were investigated to reveal novel insights into their genetic diversity. In particular, in the highly repetitive genome of rye the genome complexity reduction achieved by RNA-seq is beneficial. The comprehensive investigation of the rye transcriptome provided new information of its genetic diversity. Therefore, the first rye transcriptome reference was constructed and utilized for variant discovery. The study revealed ~18,000 single nucleotide variants (SNVs) in coding regions. With the subsequent design of a genotyping assay (RYE5k) this knowledge was successfully converted into a resource for application in breeding programs. The identification of genomic variants requires a high degree of accuracy. Two methods were developed that assist to increase the accuracy in the process of variant discovery: the 'combinatorial variant calling' and the approach of '*k*-mer repeat investigation'. With the first method, the reliability of variant calling was increased by the interlaced support and analysis of multiple detection procedures. Successfulness of the approach was shown by determining the diversity in biomass-related genes of maize. Hereby, the applied capture sequencing approach revealed 86,875 SNVs in coding regions. The second method was motivated by the complexity of the large and repetitive barley genome. An in-depth survey of repeats facilitated to improve diversity detection. Therefore, *k*-mer analyses were used to gain knowledge of repetitive features and this resulted in greater precision in the subsequent variant calling. This positive effect was shown in a genome-wide diversity study of barley, where a large proportion of variant positions were discarded because of ambiguous repeat sequences. As a result, more than 15 million high-quality SNVs were identified in five diverse barley cultivars and an additional accession of the wild progenitor of cultivated barley. The study successfully revealed novel and genome wide insights into the genetic diversity of barley.

All evaluated DNA sequencing concepts were shown to perform effectively in diversity studies. As a result of this comprehensive evaluation, three considerable diversity resources were constructed for rye, maize, and barley, which will significantly assist breeding initiatives and plant science. The developed methods for improved accuracy of variant prediction were successfully applied in the challenging context of repetitive plant genomes. Beyond this, both approaches are applicable to sequences of virtually any species and have the benefit of little or no additional costs.

7. Zusammenfassung

In den vergangenen Jahren hat sich die Hochdurchsatz-Sequenzierung als revolutionäre Technologie etabliert. Seitdem können selbst große und komplexe Pflanzengenome entschlüsselt werden. Um die genetische Information einer Spezies zu sequenzieren, sind drei wesentliche Strategien zu unterscheiden: die Verfahren der Transkriptom- (RNA-Seq), der Target Capture- (CAP-Seq) und der Whole-Genome-Shotgun- (WGS-Seq) Sequenzierung. Einen Vergleich der Anwendbarkeit in komplexen Genomen gab es bisher nicht. In dieser Doktorarbeit wurden die Verfahren auf ihr Potential zur Detektion genetischer Diversität evaluiert und Methoden entwickelt, welche die Vorhersagegenauigkeit verbessern.

Die Kulturpflanzen Roggen (*Secale cereale* L.), Mais (*Zea mays* L.) und Gerste (*Hordeum vulgare* L.) sind hierfür untersucht worden, um bisher nicht bekannte genetische Diversitätsmerkmale zu erhalten. Insbesondere in hochrepetitiven Genomen, wie dem des Roggens, bietet die Komplexitätsreduktion der Transkriptom-Sequenzierung eine wichtige Herangehensweise. In diesem Projekt wurden wesentliche Resultate zur genetischen Diversität des Roggens erzeugt. Aus der dabei etablierten Transkriptom-Referenzsequenz sind ~18.000 Einzelnukleotid-Variationen (SNVs) in genkodierenden Sequenzbereichen detektiert worden. Mit der Konstruktion eines Genotypisierungs-Arrays (RYE5k) wurden diese erfolgreich als Ressource für die Züchtungsforschung bereitgestellt. Essentiell für die Detektion von Variationen auf Nukleotidebene ist eine hohe Vorhersagegenauigkeit. Dafür sind zwei Methoden zur Verbesserung entwickelt worden: das Verfahren der „kombinatorischen Diversitätsdetektion“ und die „*K*-mer Repeat Analyse“. In der ersten Methode konnte durch eine vernetzte Analyse mehrerer Detektionsverfahren die Zuverlässigkeit der Vorhersage weiter erhöht werden. Das Verfahren wurde erfolgreich in einer Studie angewendet, in der die Diversität von Kandidatengenen in Mais untersucht wurde, welche mit Biomasse im Zusammenhang stehen. Unter Verwendung der zielgerichteten Target-Sequenzierung sind 86.875 SNVs in genkodierenden Sequenzbereichen identifiziert worden. Die Entwicklung der zweiten Methode war durch die Komplexität des hochrepetitiven Gerstengenoms motiviert. Um eine verbesserte Diversitätserkennung zu etablieren, ist zusätzlich eine *k*-mer Analyse für die detektierten Sequenzvariationen angewendet worden. Mit den erhaltenen Informationen zu Repeatmustern wurde die Vorhersagegenauigkeit verbessert. Für diese Gerstendiversitätsstudie wurde die vollständige Genomsequenz von fünf Gerstensorten und einer Wildgerste mittels WGS-Sequenzierung analysiert. Die durchgeführten Analysen konnten belegen, dass ein Großteil der initial detektierten

Sequenzvariationen aufgrund kritischer Repeatanteile fehlerhaft war und infolgedessen herausgefiltert werden musste. Im Ergebnis konnten über 15 Millionen qualitative SNVs identifiziert und damit ein erster genomweiter Eindruck der genetischen Diversität der Gerste gewonnen werden.

Alle in dieser Arbeit evaluierten Sequenzierungsstrategien sind erfolgreich in Diversitätsstudien angewendet worden. Mit den Ergebnissen dieser Evaluierung und den darin erbrachten Analysen zur genetischen Diversität konnten wichtige Ressourcen für Roggen, Mais und Gerste etabliert werden. Diese können auf zukünftige Züchtungsprogramme und auf die Pflanzenforschung einen maßgeblichen Einfluss haben. Die entwickelten Verfahren zur verbesserten Diversitätsdetektion wurden erfolgreich in komplexen Pflanzengenomen angewendet, sind speziesübergreifend nutzbar und mit geringen zusätzlichen Kosten verbunden.

8. Bibliography

1. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, et al. Mining SNPs From EST Databases. Genome Res. 1999;9: 167–174. doi:10.1101/gr.9.2.167

2. CBD. Year in Review 2010. 2010.

3. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326: 1112–1115. doi:10.1126/science.1178534

4. Varshney RK, Marcel TC, Ramsay L, Russell J, Röder MS, Stein N, et al. A high density barley microsatellite consensus map with 775 SSR loci. Theor Appl Genet. 2007;114: 1091–1103. doi:10.1007/s00122-007-0503-7

5. Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, et al. Development and implementation of high-throughput SNP genotyping in barley. 2009;13: 1–13. doi:10.1186/1471-2164-10-582

6. Stein N, Prasad M, Scholz U, Thiel T, Zhang H, Wolf M, et al. A 1,000-loci transcript map of the barley genome: New anchoring points for integrative grass genomics. Theor Appl Genet. 2007;114: 823–839. doi:10.1007/s00122-006-0480-2

7. Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, et al. De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. BMC Genomics. 2009;10: 547. doi:10.1186/1471-2164-10-547

8. Mayer KFX, Taudien S, Martis M, Simková H, Suchánková P, Gundlach H, et al. Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 2009;151: 496–505. doi:10.1104/pp.109.142612

9. Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, et al. The international barley sequencing consortium--at the threshold of efficient access to the barley genome. Plant Physiol. 2009;149: 142–147. doi:10.1104/pp.108.128967

10. Bartos J, Paux E, Kofler R, Havránková M, Kopecký D, Suchánková P, et al. A first survey of the rye (Secale cereale) genome composition through BAC end sequencing of the short arm of chromosome 1R. BMC Plant Biol. 2008;8: 95. doi:10.1186/1471-2229-8-95

11. Mardis E, McPherson JD, Martienssen R, Wilson RK, McCombie WR. What is finished, and why does it matter? Genome Res. 2002;12: 669–671. doi:10.1101/gr.032102.O

12. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing : lessons and rationales. Trends Plant Sci. Elsevier Ltd; 2011;16: 77–88. doi:10.1016/j.tplants.2010.10.005

13. Sanger F, Coulson AR. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. J Mol Biol. 1975;94: 441–448.

14. Maxam a M, Gilbert W. A new method for sequencing DNA. 1977. Biotechnology. 1992;24: 99–103. Available: http://www.ncbi.nlm.nih.gov/pubmed/1422074

15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. PNAS. 1977;74: 5463–5467.

16. Lloyd M. Smith, Jane Z. Sanders, Robert J. Kaiser, Peter Hughes, Chris Dodd, Charles R. Connell, Cheryl Heiner SBHK& LEH. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321: 674–679. doi:doi:10.1038/321674a0

17. IHGSC. Finishing the euchromatic sequence of the human genome. Nature. 2004;431: 931–945. doi:10.1038/nature03001

18. IHGSC. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921. doi:10.1038/35057062

19. Schloss JA. How to get genomes at one ten-thousandth the cost. Nat Biotechnol. 2008;26: 1113–1115.

20. Merriman B, R&D Team IT, Rothberg JM. Progress in Ion Torrent semiconductor chip based sequencing. Electrophoresis. 2012;33: 3397–3417. doi:10.1002/elps.201200424

21. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. Nature Publishing Group; 2011;475: 348–352. doi:10.1038/nature10242

22. Munroe DJ, Harris TJR. Third-generation sequencing fireworks at Marco Island. Nat Biotechnol. Nature Publishing Group; 2010;28: 426–8. doi:10.1038/nbt0510-426

23. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of Next-Generation Sequencing Technologies. Anal Chem. 2011;83: 4327–4341. doi:10.1021/ac2010857.Landscape

24. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat Biotechnol. 2008;26: 1117–1124. doi:10.1038/nbt1485

25. Ronaghi M. Pyrosequencing Sheds Light on DNA Sequencing Pyrosequencing Sheds Light on DNA Sequencing. Genome Res. 2001;11: 3–11. doi:10.1101/gr.150601

26. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. 2005;437: 376–381. doi:10.1038/nature03959

27. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012: 251364. doi:10.1155/2012/251364

28. Roche Diagnostics GmbH. GS FLX+ System: Sanger-like read lengths - the power of next-gen throughput [Internet]. 2011. Available: http://454.com/downloads/GSFLXApplicationFlyer_FINALv2.pdf

29. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59. doi:10.1038/nature07517

30. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. Nature Publishing Group; 2010;11: 31–46. doi:10.1038/nrg2626

31. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics. BMC Genomics; 2012;13: 1. doi:10.1186/1471-2164-13-341

32. Laehnemann D, Borkhardt a., McHardy a. C. Denoising DNA deep sequencing data--high-throughput sequencing errors and their correction. Brief Bioinform. 2015; 1–26. doi:10.1093/bib/bbv029

33. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19: 227–240. doi:10.1093/hmg/ddq416

34. Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics. 2014;15: 699. doi:10.1186/1471-2164-15-699

35. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One. 2012;7: 1–12. doi:10.1371/journal.pone.0047768

36. Liao Y-C, Lin S-H, Lin H-H. Completing bacterial genome assemblies: strategy and performance comparisons. Sci Rep. 2015;5: 8747. doi:10.1038/srep08747

37. Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. Bioessays. 2010;32: 524–36. doi:10.1002/bies.200900181

38. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30. doi:10.1016/j.tig.2014.07.001

39. Thudi M, Li Y, Jackson S a, May GD, Varshney RK. Current state-of-art of sequencing technologies for plant genomics research. Brief Funct Genomics. 2012;11: 3–11. doi:10.1093/bfgp/elr045

40. Varshney RK, Terauchi R, Mccouch SR. Harvesting the Promising Fruits of Genomics : Applying Genome Sequencing Technologies to Crop Breeding. PLOS Biol. 2014;12. doi:10.1371/journal.pbio.1001883

41. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. Nature Publishing Group; 2011;470: 198–203. doi:10.1038/nature09796

42. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10: R32. doi:10.1186/gb-2009-10-3-r32

43. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15: 121–32. doi:10.1038/nrg3642

44. Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, et al. NIH Public Access. 2013; doi:10.4172/2153-0602.1000136.A

45. Archer J, Baillie G, Watson SJ, Kellam P, Rambaut A, Robertson DL. Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics. BioMed Central Ltd; 2012;13: 47. doi:10.1186/1471-2105-13-47

46. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in C. elegans. Nat Methods. 2008;5: 183–8. doi:10.1038/nmeth.1179

47. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. BioMed Central Ltd; 2011;12: R112. doi:10.1186/gb-2011-12-11-r112

48. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40: e72. doi:10.1093/nar/gks001

49. Yang X, Chockalingam SP, Aluru S. A survey of error-correction methods for next-generation sequencing. Brief Bioinform. 2013;14: 56–66. doi:10.1093/bib/bbs015

50. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9: 517. doi:10.1186/1471-2164-9-517

51. El-Metwally S, Hamza T, Zakaria M, Helmy M. Next-generation sequence assembly: four stages of data processing and computational challenges. PLoS Comput Biol. 2013;9: e1003345. doi:10.1371/journal.pcbi.1003345

52. Edwards D, Batley J, Snowdon RJ. Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet. 2013;126: 1–11. doi:10.1007/s00122-012-1964-x

53. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408: 796–815. doi:doi:10.1038/35048692

54. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, et al. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. Nature. Nature Publishing Group; 2013;496: 91–5. doi:10.1038/nature12028

55. Michael TP, Jackson S. The First 50 Plant Genomes. Plant Genome. 2013; doi:10.3835/plantgenome2013.03.0001in

56. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al. The Norway spruce genome sequence and conifer genome evolution. Nature. Nature Publishing Group; 2013;497: 579–84. doi:10.1038/nature12211

57. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43: 956–63. doi:10.1038/ng.911

58. Hamilton JP, Buell CR. Advances in plant genome sequencing. Plant J. 2012;70: 177–90. doi:10.1111/j.1365-313X.2012.04894.x

59. Banks J a, Et-Al. The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science. 2011;332: 960–963. doi:10.1126/science.1203810

60. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. Nature Publishing Group; 2009;462: 315–22. doi:10.1038/nature08514

61. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. Nature Publishing Group; 2014; 1–15. doi:10.1038/nmeth.3065

62. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. Elsevier Inc.; 2011;147: 1408–19. doi:10.1016/j.cell.2011.11.013

63. Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. Nat Protoc. Nature Publishing Group; 2013;8: 539–54. doi:10.1038/nprot.2013.023

64. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. Nat Methods. Nature Publishing Group; 2014; doi:10.1038/nmeth.3104

65. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484

66. Weber M, Horn F, Sieber P, Hellwig D, Riege K, Marz M, et al. Defining the transcriptomic landscape of Candida glabrata by RNA-Seq J org. 2015;43: 1392–1406. doi:10.1093/nar/gku1357

67. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9: e78644. doi:10.1371/journal.pone.0078644

68. Robles J a, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC Genomics. 2012;13: 484. doi:10.1186/1471-2164-13-484

69. Hou Z, Jiang P, Swanson S a., Elwell AL, Nguyen BKS, Bolin JM, et al. A cost-effective RNA sequencing protocol for large-scale gene expression studies. Sci Rep. 2015;5: 9570. doi:10.1038/srep09570

70. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37. doi:10.1093/nar/gkp596

71. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. Nature Publishing Group; 2011;12: 87–98. doi:10.1038/nrg2934

72. Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. Nucleic Acids Res. 2007;35. doi:10.1093/nar/gkm683

73. Zhang X, Rosen BD, Tang H, Krishnakumar V, Town CD. Polyribosomal RNA-Seq Reveals the Decreased Complexity and Diversity of the Arabidopsis Translatome. PLoS One. 2015;10: e0117699. doi:10.1371/journal.pone.0117699

74. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. Plant J. 2007;51: 910–8. doi:10.1111/j.1365-313X.2007.03193.x

75. Oliver RE, Lazo GR, Lutz JD, Rubenfield MJ, Tinker NA, Anderson JM, et al. Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. BMC Genomics. BioMed Central Ltd; 2011;12: 77. doi:10.1186/1471-2164-12-77

76. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next- generation sequencing. Nat Methods. Nature Publishing Group; 2010;7: 111–118. doi:10.1038/nmeth.1419

77. Mascher M, Richmond T a, Gerhardt DJ, Himmelbach A, Clissold L, Sampath D, et al. Barley whole exome capture : a tool for genomic research in the genus Hordeum and beyond. Plant J. 2013;76: 494–505. doi:10.1111/tpj.12294

78. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. Nature Publishing Group; 2007;39: 1522–1527. Available: http://dx.doi.org/10.1038/ng.2007.42

79. Neves LG, Davis JM, Barbazuk WB, Kirst M. Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J. 2013;75: 146–156. doi:10.1111/tpj.12193

80. Saintenac C, Jiang D, Wang S, Akhunov E. Sequence-based mapping of the polyploid wheat genome. G3 (Bethesda). 2013;3: 1105–14. doi:10.1534/g3.113.005819

81. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457: 551–6. doi:10.1038/nature07723

82. Li J-Y, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience. 2014;3: 8. doi:10.1186/2047-217X-3-8

83. Zimin a., Stevens K a., Crepeau MW, Holtz-Morris a., Koriabine M, Marcais G, et al. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics. 2014;196: 875–890. doi:10.1534/genetics.113.159715

84. Neale DB, Wegrzyn JL, Stevens K a, Zimin A V, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15: R59. doi:10.1186/gb-2014-15-3-r59

85. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, et al. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11: 25–37. doi:10.1093/bfgp/elr035

86. Bonfield JK, Smith KF, Staden R. A new DNA sequence assembly program. Nucleic Acids Res. 1995;23: 4992–4999.

87. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9: 868–877. doi:10.1101/gr.9.9.868

88. Compeau PEC, Pevzner P a, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. Nature Publishing Group; 2011;29: 987–91. doi:10.1038/nbt.2023

89. Pevzner P a, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

90. Schatz MC, Witkowski J, McCombie WR. Current challenges in de novo plant genome sequencing and assembly. Genome Biol. 2012;13: 243. doi:10.1186/gb4015

91. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11: 473–483. doi:10.1093/bib/bbq015

92. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8: 186–94. Available: http://www.ncbi.nlm.nih.gov/pubmed/9521922

93. Tilman D, Lehman C. Human-caused environmental change : Impacts on plant diversity and evolution. Proc Natl Acad Sci. 2001;98: 5433–5440. doi:10.1073/pnas.091093198

94. Brookes AJ. The essence of SNPs. Gene. 1999;234: 177–186. doi:http://dx.doi.org/10.1016/S0378-1119(99)00219-X

95. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 2007;3: 1745–1756. doi:10.1371/journal.pgen.0030163

96. Guo L, Gao Z, Qian Q. Application of resequencing to rice genomics, functional genomics and evolutionary analysis. Rice (N Y). 2014;7: 4. doi:10.1186/s12284-014-0004-7

97. Sim S-C, Durstewitz G, Plieske J, Wieseke R, Ganal MW, Van Deynze A, et al. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. PLoS One. 2012;7: e40563. doi:10.1371/journal.pone.0040563

98. Carr M, Cotton S, Rogers DW, Pomiankowski A, Smith H, Fowler K. Assigning sex to pre-adult stalk-eyed flies using genital disc morphology and X chromosome zygosity. BMC Dev Biol. 2006;6: 29. doi:10.1186/1471-213X-6-29

99. Khalak HG, Wakil SM, Imtiaz F, Ramzan K, Baz B, Almostafa A, et al. Autozygome maps dispensable DNA and reveals potential selective bias against nullizygosity. Genet Med. 2012;14: 515–519. doi:10.1038/gim.2011.28

100. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330: 641–646. doi:10.1126/science.1197005

101. Foxe JP, Stift M, Tedder A, Haudry A, Wright SI, Mable BK. Reconstructing origins of loss of self-incompatibility and selfing in North American arabidopsis lyrata: A population genetic context. Evolution (N Y). 2010;64: 3495–3510. doi:10.1111/j.1558-5646.2010.01094.x

102. Schoen DJ, Brown a H. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proc Natl Acad Sci U S A. 1991;88: 4494–4497. doi:10.1073/pnas.88.10.4494

103. Mable BK, Adam a. Patterns of genetic diversity in outcrossing and selfing populations of Arabidopsis lyrata. Mol Ecol. 2007;16: 3565–3580. doi:10.1111/j.1365-294X.2007.03416.x

104. Baumann U. Self-incompatibility in the Grasses. Ann Bot. 2000;85: 203–209. doi:10.1006/anbo.1999.1056

105. Morrell PL, Toleno DM, Lundy KE, Clegg MT. Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. PNAS. 2005;102: 2442–2447. doi:10.1073/pnas.0409804102

106. Koelling V a, Hamrick JL, Mauricio R. Genetic diversity and structure in two species of Leavenworthia with self-incompatible and self-compatible populations. Heredity (Edinb). Nature Publishing Group; 2011;106: 310–318. doi:10.1038/hdy.2010.59

107. Barrett SCH. The evolution of plant sexual diversity. Nat Rev Genet. 2002;3: 274–284. doi:10.1038/nrg776

108. Ness RW, Wright SI, Barrett SCH. Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylous plant Eichhornia paniculata. Genetics. 2010;184: 381–392. doi:10.1534/genetics.109.110130

109. Nybom H. Comparison of different nuclear DNA markers for estimating intraspecific genetic diversity in plants. Mol Ecol. 2004;13: 1143–1155. doi:10.1111/j.1365-294X.2004.02141.x

110. Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, et al. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? PLoS Biol. 2012;10. doi:10.1371/journal.pbio.1001388

111. Falk DE. Generating and maintaining diversity at the elite level in crop breeding. Genome. 2010;53: 982–991. doi:10.1139/G10-081

112. FAO. How to Feed the World in 2050. Proceedings of a Technical Meeting of Experts (FAO, Rome). 2009. pp. 1–35. doi:10.1111/j.1728-4457.2009.00312.x

113. Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. Plant genome sequencing — applications for crop improvement. Curr Opin Biotechnol. 2014;26: 31–37. doi:10.1016/j.copbio.2013.08.019

114. Kuczyńska A, Surma M, Adamski T, Mikołajczak K, Krystkowiak K, Ogrodowicz P. Effects of the semi-dwarfing sdw1/denso gene in barley. J Appl Genet. 2013;54: 381–390. doi:10.1007/s13353-013-0165-x

115. Strimbu K, Tavel J a. What are Biomarkers? Curr Opin HIV AIDS. 2011;5: 463–466. doi:10.1097/COH.0b013e32833ed177

116. Maltais K, Houde M. A new biochemical marker for aluminium tolerance in plants. Physiol Plant. Blackwell Science, Ltd; 2002;115: 81–86. doi:10.1034/j.1399-3054.2002.1150109.x

117. Kwiatek M, Wiśniewska H, Apolinarska B. Cytogenetic analysis of Aegilops chromosomes, potentially usable in triticale (X Triticosecale Witt.) breeding. J Appl Genet. 2013;54: 147–155. doi:10.1007/s13353-013-0133-5

118. Delibes A, Otero C, García-Olmedo F, Dosba F. Biochemical markers associated with two Mv chromosomes from Aegilops ventricosa in wheat-Aegilops addition lines. Theor Appl Genet. Springer-Verlag; 1981;60: 5–10. doi:10.1007/BF00275171

119. Botstein D, White RL, Skolnick M, Davis RW. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. 1980; 314–331.

120. Sahu BB, Sumit R, Srivastava SK, Bhattacharyya MK. Sequence based polymorphic (SBP) marker technology for targeted genomic regions: its application in generating a molecular map of the Arabidopsis thaliana genome. BMC Genomics. BioMed Central Ltd; 2012;13: 20. doi:10.1186/1471-2164-13-20

121. Lateef DD. DNA Marker Technologies in Plants and Applications for Crop Improvements. J Biosci Med. 2015;3: 7–18. doi:http://dx.doi.org/10.4236/jbm.2015.35002

122. Semagn K, Bjornstad a., Ndjiondjop MN. An overview of molecular marker methods for plants. African J Biotechnol. 2006;5: 2540–2568. doi:10.1111/j.1439-0523.2009.01731.x

123. Vos P, Hogers R, Bleeker M, Reijans M, Lee T Van De, Frijters A, et al. AFLP : a new technique for DNA fingerprinting. Nucleic Acids Res. 1995;23: 4407–4414. doi:10.1093/nar/23.21.4407

124. Baird N a, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis Z a, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3: e3376. doi:10.1371/journal.pone.0003376

125. Elshire RJ, Glaubitz JC, Sun Q, Poland J a, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6: e19379. doi:10.1371/journal.pone.0019379

126. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS One. 2013;8: 1–9. doi:10.1371/journal.pone.0054603

127. Blair MW, Cortés AJ, Penmetsa RV, Farmer A, Carrasquilla-Garcia N, Cook DR. A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (Phaseolus vulgaris L.). Theor Appl Genet. 2013;126: 535–48. doi:10.1007/s00122-012-1999-z

128. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, et al. A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS One. 2011;6: e28334. doi:10.1371/journal.pone.0028334

129. Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Booth A, et al. Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Genet Genomics. 2005;274: 515–27. doi:10.1007/s00438-005-0046-z

130. Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, et al. Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J. 2012;10: 743–9. doi:10.1111/j.1467-7652.2012.00718.x

131. Syvänen A-C. Toward genome-wide SNP genotyping. Nat Genet. 2005;37 Suppl: S5–S10. doi:10.1038/ng1558

132. Gray IC, Campbell D a, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet. 2000;9: 2403–2408. doi:10.1093/hmg/9.16.2403

133. Campbell WH, Gowri G. Codon usage in higher plants, green algae, and cyanobacteria. Plant Physiol. 1990;92: 1–11. doi:10.1104/pp.92.1.1

134. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 80–92.

135. Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, et al. Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS One. 2012;7: e31745. doi:10.1371/journal.pone.0031745

136. Yan J, Yang X, Shah T, Sánchez-Villeda H, Li J, Warburton M, et al. High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed. 2009;25: 441–451. doi:10.1007/s11032-009-9343-2

137. Kilian B, Graner A. NGS technologies for analyzing germplasm diversity in genebanks. Brief Funct Genomics. 2012;11: 38–50. doi:10.1093/bfgp/elr046

138. Chen H, He H, Zou Y, Chen W, Yu R, Liu X, et al. Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (Oryza sativa L.). Theor Appl Genet. 2011;123: 869–79. doi:10.1007/s00122-011-1633-5

139. Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, et al. Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. PLoS One. 2013;8: e84136. doi:10.1371/journal.pone.0084136

140. Gudbjartsson DF, Helgason H, Gudjonsson S a, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015; doi:10.1038/ng.3247

141. Abecasis GR, Auton A, Brooks LD, DePristo M a, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491: 56–65. doi:10.1038/nature11632

142. Takeda S, Matsuoka M. Genetic approaches to crop improvement: responding to environmental and population changes. Nat Rev Genet. 2008;9: 444–57. doi:10.1038/nrg2342

143. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs R a, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–73. doi:10.1038/nature09534

144. Danecek P, Auton A, Abecasis G, Albers C a, Banks E, DePristo M a, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–8. doi:10.1093/bioinformatics/btr330

145. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009;10: R134. doi:10.1186/gb-2009-10-11-r134

146. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. BioMed Central Ltd; 2013;5: 28. doi:10.1186/gm432

147. Li J-W, Schmieder R, Ward RM, Delenick J, Olivares EC, Mittelman D. SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics. 2012;28: 1272–3. doi:10.1093/bioinformatics/bts128

148. Kumar S, Banks TW, Cloutier S. SNP Discovery through Next-Generation Sequencing and Its Applications. Int J Plant Genomics. 2012;2012: 831460. doi:10.1155/2012/831460

149. Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 2002;5: 94–100. doi:http://dx.doi.org/10.1016/S1369-5266(02)00240-6

150. Cheng AY, Teo YY, Ong RTH. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics. 2014;30: 1707–1713. doi:10.1093/bioinformatics/btu067

151. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2013; doi:10.1093/bib/bbs086

152. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014;15: 244. doi:10.1186/1471-2164-15-244

153. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

154. Li H. Improving SNP discovery by base alignment quality. Bioinformatics. 2011;27: 1157–1158. doi:10.1093/bioinformatics/btr076

155. Xu F, Wang W, Wang P, Jun Li M, Chung Sham P, Wang J. A fast and accurate SNP detection algorithm for next-generation sequencing data. Nat Commun. Nature Publishing Group; 2012;3: 1258. doi:10.1038/ncomms2256

156. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25: 2283–2285. doi:10.1093/bioinformatics/btp373

157. Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Boroevich K a, et al. A practical method to detect SNVs and indels from whole genome and exome sequencing data. Sci Rep. 2013;3: 2161. doi:10.1038/srep02161

158. Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 2010;26: i318–24. doi:10.1093/bioinformatics/btq214

159. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39: e132. doi:10.1093/nar/gkr599

160. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, et al. A general approach to single-nucleotide polymorphism discovery. Nat Genet. 1999;23: 452–456.

161. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012; 1–9.

162. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat Publ Gr. Nature Publishing Group; 2014;15: 749–763. doi:10.1038/nrg3803

163. You N, Murillo G, Su X, Zeng X, Xu J, Ning K, et al. SNP calling using genotype model selection on high-throughput sequencing data. Bioinformatics. 2012;28: 643–650. doi:10.1093/bioinformatics/bts001

164. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19: 1124–32. doi:10.1101/gr.088013.108

165. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, et al. Articles Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Publ Gr. Nature Publishing Group; 2010;42: 931–936. doi:10.1038/ng.691

166. Martin J a, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. Nature Publishing Group; 2011;12: 671–82. doi:10.1038/nrg3068

167. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature. 2000;407: 513–516. doi:10.1038/35035083

168. DePristo M a, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–8. doi:10.1038/ng.806

169. Zharag J, Wheeler D a., Yakub I, Wei S, Sood R, Rowe W, et al. SNPdetector: A software tool for sensitive accurate SNP detection. PLoS Comput Biol. 2005;1: 0395–0404. doi:10.1371 /journal.pcbi.0010053

170. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. Nature Publishing Group; 2012;13: 36–46. doi:10.1038/nrg3117

171. Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. Nat Biotechnol. Nature Publishing Group; 2012;30: 1084–94. doi:10.1038/nbt.2421

172. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res. 2010;20: 1165–73. doi:10.1101/gr.101360.109

173. IBSC. A physical, genetic and functional sequence assembly of the barley genome. Nature. Nature Publishing Group; 2012;491: 711–6. doi:10.1038/nature11543

174. Duarte J, Rivière N, Baranger A, Aubert G, Burstin J, Cornet L, et al. Transcriptome sequencing for high throughput SNP development and genetic mapping in Pea. BMC Genomics. 2014;15: 126. doi:10.1186/1471-2164-15-126

175. Szostak E, Gebauer F. Translational control by 3'-UTR-binding proteins. Brief Funct Genomics. 2013;12: 58–65. doi:10.1093/bfgp/els056

176. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol. 2002;3: REVIEWS0004. doi:10.1186/gb-2002-3-3-reviews0004

177. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19: 1068–76. doi:10.1101/gr.089516.108

178. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8: 61–65. doi:10.1038/nmeTh.1527

179. Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, van Houdt J, Maes GE, et al. SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (Clupea harengus). PLoS One. 2012;7: e42089. doi:10.1371/journal.pone.0042089

180. Bevan MW, Uauy C. Genomics reveals new landscapes for crop improvement. Genome Biol. 2013;14: 206. doi:10.1186/gb-2013-14-6-206

181. Bodi K, Perera a G, Adams PS, Bintzler D, Dewar K, Grove DS, et al. Comparison of commercially available target enrichment methods for next-generation sequencing. J Biomol Tech. 2013;24: 73–86. doi:10.7171/jbt.13-2402-002

182. Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, et al. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat Protoc. 2014;9: 989–1009. doi:10.1038/nprot.2014.058

183. Chen Y, Liu T, Yu C, Chiang T, Hwang C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. 2013;8. doi:10.1371/journal.pone.0062856

184. Russell J, Booth A, Fuller J, Harrower B, Hedley P, Machray G, et al. A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome. 2004;47: 389–398. doi:10.1139/G03-125

185. Pellino M, Hojsgaard D, Schmutzer T, Scholz U, Hörandl E, Vogel H, et al. Asexual genome evolution in the apomictic Ranunculus auricomus complex: examining the effects of hybridization and mutation accumulation. Mol Ecol. 2013; doi:10.1111/mec.12533

186. Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, et al. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. Proc Natl Acad Sci U S A. 2012;109: 13343–6. doi:10.1073/pnas.1204237109

187. Kono TJY, Seth K, Poland J a., Morrell PL. SNPMeta: SNP annotation and SNP metadata collection without a reference genome. Mol Ecol Resour. 2014;14: 419–425. doi:10.1111/1755-0998.12183

188. ORF Finder [Internet]. Available: http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi

189. Min XJ, Butler G, Storms R, Tsang A. OrfPredictor: Predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 2005;33: 677–680. doi:10.1093/nar/gki394

190. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. Elsevier; 2015;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

191. Marth GT, Czabarka E, Murvai J, Sherry ST. The Allele Frequency Spectrum in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in Three Large World Populations. Genetics. 2004;166: 351–372. doi:10.1534/genetics.166.1.351

192. Romay MC, Millard MJ, Glaubitz JC, Peiffer J a, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. BioMed Central Ltd; 2013;14: R55. doi:10.1186/gb-2013-14-6-r55

193. Muraya MM, Schmutzer T, Ulpinnis C, Scholz U, Altmann T. Targeted Sequencing Reveals Large-Scale Sequence Polymorphism in Maize Candidate Genes for Biomass Production and Composition. PLoS One. 2015;10: e0132120. doi:10.1371/journal.pone.0132120

194. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. Nature Publishing Group; 2012;44: 803–7. doi:10.1038/ng.2313

195. Muñoz-Amatriaín M, Cuesta-Marcos A, Endelman JB, Comadran J, Bonman JM, Bockelman HE, et al. The USDA barley core collection: genetic diversity, population structure, and potential for genome-wide association studies. PLoS One. 2014;9: e94688. doi:10.1371/journal.pone.0094688

196. Frascaroli E, Schrag T a, Melchinger AE. Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. Theor Appl Genet. 2013;126: 133–41. doi:10.1007/s00122-012-1968-6

197. Mascher M, Wu S, Amand PS, Stein N, Poland J. Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms : A Comparison of Genetic and Reference-Based Marker Ordering in Barley. 2013;8: 1–11. doi:10.1371/journal.pone.0076925

198. Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, et al. How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min. 2012;5: 6. doi:10.1186/1756-0381-5-6

199. Li H. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM. 2013;00: 1–3.

200. Lunter G, Goodson M. Stampy : A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. 2011; 936–939. doi:10.1101/gr.111120.110.tions

201. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, et al. RNA sequencing reveals the complex regulatory network in the maize kernel. Nat Commun. Nature Publishing Group; 2013;4: 2832. doi:10.1038/ncomms3832

202. Comadran J, Kilian B, Russell J, Ramsay L, Stein N, Ganal M, et al. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat Genet. Nature Publishing Group; 2012;44: 1388–92. doi:10.1038/ng.2447

203. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLoS One. 2012;7: 1–6. doi:10.1371/journal.pone.0052249

204. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski E a, et al. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 2010;20: 273–80. doi:10.1101/gr.096388.109

205. Wu C, Cui Y. Boosting signals in gene-based association studies via efficient SNP selection. Brief Bioinform. 2014;15: 279–91. doi:10.1093/bib/bbs087

206. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA. Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Mol Plant. Elsevier Ltd; 2015;8: 831–846. doi:10.1016/j.molp.2015.02.002

207. Li H. Exploring single-sample snp and indel calling with whole-genome de novo assembly. Bioinformatics. 2012;28: 1838–1844. doi:10.1093/bioinformatics/bts280

208. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 2002;3: 329–341. Available: http://dx.doi.org/10.1038/nrg793

209. Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004;32: D360–D363. doi:10.1093/nar/gkh099

210. Flavell RB, Bennett MD, Smith JB, Smith DB. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet. 1974;12: 257–269. doi:10.1007/BF00485947

211. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30: 2843–2851. doi:10.1093/bioinformatics/btu356

212. Wicker T, Matthews DE, Keller B. TREP: a database for Triticeae repetitive elements. Trends Plant Sci. 2002;7: 561–562.

213. Smit A, Hubley R, Green P. RepeatMasker [Internet]. 1996. doi:http://www.repeatmasker.org

214. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol. 2006;13: 1028–1040. doi:10.1089/cmb.2006.13.1028

215. Koch P, Platzer M, Downie BR. RepARK - de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 2014;42: 1–12. doi:10.1093/nar/gku210

216. Chapman J a, Mascher M, Buluç A, Barry K, Georganas E, Session A, et al. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 2015;16: 1–17. doi:10.1186/s13059-015-0582-8

217. Chor B, Horn D, Goldman N, Levy Y, Massingham T. Genomic DNA k-mer spectra: models and modalities. Genome Biol. 2009;10: R108. doi:10.1186/gb-2009-10-10-r108

218. Wicker T, Narechania A, Sabot F, Stein J, Vu GTH, Graner A, et al. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. BMC Genomics. 2008;9: 518. doi:10.1186/1471-2164-9-518

219. Gurushidze M, Hensel G, Hiekel S, Schedel S, Valkov V, Kumlehn J. True-breeding targeted gene knock-out in barley using designer TALE-nuclease in haploid cells. PLoS One. 2014;9: e92046. doi:10.1371/journal.pone.0092046

220. Salzberg SL, Yorke J a. Beware of mis-assembled genomes. Bioinformatics. 2005;21: 4320–4321. doi:10.1093/bioinformatics/bti769

221. Kelley DR, Salzberg SL. Detection and correction of false segmental duplications caused by genome mis-assembly. Genome Biol. 2010;11: R28. doi:10.1186/gb-2010-11-3-r28

222. Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. Nature Publishing Group; 2014;32. doi:10.1038/nbt.2979

223. Deokar A a, Ramsay L, Sharpe AG, Diapari M, Sindhu A, Bett K, et al. Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. BMC Genomics. 2014;15: 708. doi:10.1186/1471-2164-15-708

224. Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Buell CR. Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. PLoS One. 2012;7: e33071. doi:10.1371/journal.pone.0033071

225. Yi M, Zhao Y, Jia L, He M, Kebebew E, Stephens RM. Performance comparison of SNP detection tools with illumina exome sequencing data - An assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Res. 2014;42: 1–14. doi:10.1093/nar/gku392

226. Snowdon RJ, Abbadi A, Kox T, Schmutzer T, Leckband G. Heterotic Haplotype Capture: precision breeding for hybrid performance. Trends Plant Sci. 2015;20: 410–413. doi:http://dx.doi.org/10.1016/j.tplants.2015.04.013

227. Greenfield P, Duesing K, Papanicolaou A, Bauer DC. Blue: correcting sequencing errors using consensus and context. Bioinformatics. 2014;30: 2723–2732. doi:10.1093/bioinformatics/btu368

228. Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. BioMed Central Ltd; 2010;11: R116. doi:10.1186/gb-2010-11-11-r116

229. Salmela L. Correction of sequencing errors in a mixed set of reads. Bioinformatics. 2010;26: 1284–1290. doi:10.1093/bioinformatics/btq151

230. Schröder J, Schröder H, Puglisi SJ, Sinha R, Schmidt B. SHREC: A short-read error correction method. Bioinformatics. 2009;25: 2157–2163. doi:10.1093/bioinformatics/btp379

231. Ilie L, Fazayeli F, Ilie S. HiTEC: accurate error correction in high-throughput sequencing data. Bioinformatics. 2011;27: 295–302. doi:10.1093/bioinformatics/btq653

232. Yang X, Aluru S, Dorman KS. Repeat-aware modeling and correction of short read errors. BMC Bioinformatics. BioMed Central Ltd; 2011;12: S52. doi:10.1186/1471-2105-12-S1-S52

233. Chris Ulpinnis. Evaluation von Fehlerkorrektur-Algorithmen für Sequenzdaten in komplexen Pflanzengenomen. Bachelor thesis, Martin-Luther-Universität Halle-Wittenberg, Institut für Informatik, Halle (Saale), 2013. 2013.

234. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30: 3506–3514. doi:10.1093/bioinformatics/btu538

235. Yu X, Sun S. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics. BMC Bioinformatics; 2013;14: 274. doi:10.1186/1471-2105-14-274

236. Novoalign [Internet]. 2015. Available: www.novocraft.com

237. Homer N, Nelson SF. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol. BioMed Central Ltd; 2010;11: R99. doi:10.1186/gb-2010-11-10-r99

238. Kosugi S, Natsume S, Yoshida K, MacLean D, Cano L, Kamoun S, et al. Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. PLoS One. 2013;8: e75402. doi:10.1371/journal.pone.0075402

239. Hu J, Ng PC. Predicting the effects of frameshifting indels. Genome Biol. BioMed Central Ltd; 2012;13: R9. doi:10.1186/gb-2012-13-2-r9

240. Hollister JD, Ross-Ibarra J, Gaut BS. Indel-associated mutation rate varies with mating system in flowering plants. Mol Biol Evol. 2010;27: 409–416. doi:10.1093/molbev/msp249

241. Grimm D, Hagmann J, Koenig D, Weigel D, Borgwardt K. Accurate indel prediction using paired-end short reads. BMC Genomics. 2013;14: 132. doi:10.1186/1471-2164-14-132

242. Mose LE, Wilkerson MD, Hayes DN, Perou CM, Parker JS. ABRA: improved coding indel detection via assembly-based realignment. Bioinformatics. 2014;30: 2813–2815. doi:10.1093/bioinformatics/btu376

243. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2013;15. doi:10.1093/bib/bbt069

244. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. Mol Ecol Resour. 2011;11: 123–136. doi:10.1111/j.1755-0998.2010.02943.x

245. Morin P a., Luikart G, Wayne RK. SNPs in ecology, evolution and conservation. Trends Ecol Evol. 2004;19: 208–216. doi:10.1016/j.tree.2004.01.009

246. Jansen J, de Jong AG, van Ooijen JW. Constructing dense genetic linkage maps. Theor Appl Genet. 2001;102: 1113–1122. doi:10.1007/s001220000489

247. Cheema J, Dicks J. Computational approaches and software tools for genetic linkage map estimation in plants. Brief Bioinform. 2009;10: 595–608. doi:10.1093/bib/bbp045

248. Rastas P, Paulin L, Hanski I, Lehtonen R, Auvinen P. Lep-MAP: fast and accurate linkage map construction for large SNP datasets. Bioinformatics. 2013;29: 3128–34. doi:10.1093/bioinformatics/btt563

249. Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, André I, et al. High-density SNP-based genetic map development and linkage disequilibrium assessment in Brassica napus L. 2013; 1–18. doi:10.1186/1471-2164-14-120

250. Kumar P, Gupta VK, Misra a K, Modi DR, Pandey BK. Southern Cross Journals © 2009 Potential of Molecular Markers in Plant Biotechnology. Plant Biotechnol. 2009;2: 141–162. Available: http://www.pomics.com/Pradeep_2_4_2009_141_162.pdf

251. Allen AM, Barker GL a, Wilkinson P, Burridge A, Winfield M, Coghill J, et al. Discovery and development of exome-based, co-dominant single nucleotide polymorphism markers in hexaploid wheat (Triticum aestivum L.). Plant Biotechnol J. 2013;11: 279–295. doi:10.1111/pbi.12009

252. Hiremath PJ, Kumar A, Penmetsa RV, Farmer A, Schlueter J a, Chamarthi SK, et al. Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. Plant Biotechnol J. 2012;10: 716–32. doi:10.1111/j.1467-7652.2012.00710.x

253. Wright F a, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014; doi:10.1038/ng.2951

254. Rodríguez-Ramilo ST, Toro MÁ, Caballero A, Fernández J. The accuracy of a heritability estimator using molecular information. Conserv Genet. 2007;8: 1189–1198. doi:10.1007/s10592-006-9273-z

255. Telfer EJ, Stovold GT, Li Y, Silva-Junior OB, Grattapaglia DG, Dungey HS. Parentage Reconstruction in Eucalyptus nitens Using SNPs and Microsatellite Markers: A Comparative Analysis of Marker Data Power and Robustness. PLoS One. 2015;10: e0130601. doi:10.1371/journal.pone.0130601

256. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600k SNP genotyping array. BMC Genomics. 2014;15: 823. doi:10.1186/1471-2164-15-823

257. Poland J a., Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7. doi:10.1371/journal.pone.0032253

258. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005;15: 1496–502. doi:10.1101/gr.4107905

259. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP a, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9: 356–369. doi:10.1038/nrg2344

260. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet. 2011;43: 159–62. doi:10.1038/ng.746

261. Weng J, Xie C, Hao Z, Wang J, Liu C, Li M, et al. Genome-wide association study identifies candidate genes that affect plant height in Chinese elite maize (Zea mays L.) inbred lines. PLoS One. 2011;6: e29229. doi:10.1371/journal.pone.0029229

262. Visioni A, Tondelli A, Francia E, Pswarayi A, Malosetti M, Russell J, et al. Genome-wide association mapping of frost tolerance in barley (Hordeum vulgare L.). BMC Genomics. BMC Genomics; 2013;14: 424. doi:10.1186/1471-2164-14-424

263. Nakaya A, Isobe SN. Will genomic selection be a practical method for plant breeding? Ann Bot. 2012;110: 1303–1316. doi:10.1093/aob/mcs109

264. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. SNP markers and their impact on plant breeding. Int J Plant Genomics. 2012;2012: 728398. doi:10.1155/2012/728398

265. Fischer S, Melchinger AE, Korzun V, Wilde P, Schmiedchen B, Möhring J, et al. Molecular marker assisted broadening of the Central European heterotic groups in rye with Eastern European germplasm. Theor Appl Genet. Springer-Verlag; 2010;120: 291–299. doi:10.1007/s00122-009-1124-0

266. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci. Elsevier Ltd; 2014;19: 592–601. doi:10.1016/j.tplants.2014.05.006

267. Lu Y, Shah T, Hao Z, Taba S, Zhang S, Gao S, et al. Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. PLoS One. 2011;6: e24861. doi:10.1371/journal.pone.0024861

268. Strigens A, Schipprack W, Reif JC, Melchinger AE. Unlocking the Genetic Diversity of Maize Landraces with Doubled Haploids Opens New Avenues for Breeding. PLoS One. 2013;8: 7–9. doi:10.1371/journal.pone.0057234

269. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819–1829. doi:11290733

270. Zhao Y, Mette MF, Reif JC. Genomic selection in hybrid breeding. Plant Breed. 2015;134: 1–10. doi:10.1111/pbr.12231

271. Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, Reif JC, et al. The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genomics. 2014;15: 556. doi:10.1186/1471-2164-15-556

272. Morrell PL, Buckler ES, Ross-Ibarra J. Crop genomics: advances and applications. Nat Rev Genet. Nature Publishing Group; 2011;13: 85–96. doi:10.1038/nrg3097

273. Ronald PC. Lab to Farm : Applying Research on Plant Genetics and Genomics to Crop Improvement. PLOS Biol. 2014;12. doi:10.1371/journal.pbio.1001878

274. Hayden EC. Gene sequencing leaves the laboratory Dark-matter hunt gets deep. Nature. 2013;494: 290–291.

275. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. 2011;21: 2224–41. doi:10.1101/gr.126599.111

276. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2013;2: 10. doi:10.1186/2047-217X-2-10

277. Thome OW. Flora von Deutschland Österreich und der Schweiz [Internet]. Stüber K, editor. Gera-Untermhaus: BioLib; 1903. Available: www.biolib.de

278. FAO. faostat Germany [Internet]. 2012. Available: http://faostat.fao.org/CountryProfiles/Country_Profile/Direct.aspx?lang=en&area=79

279. Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet. 2010;42: 1027–1030. doi:10.1038/ng.684

280. Bolibok-Bragoszewska H, Heller-Uszyńska K, Wenzl P, Uszyński G, Kilian A, Rakoczy-Trojanowska M. DArT markers for the rye genome - genetic diversity and mapping. BMC Genomics. 2009;10: 578. doi:10.1186/1471-2164-10-578

281. Hayden EC. The $1,000 genome. Nature. 2014;507: 295.

282. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson D a, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. Nature Publishing Group; 2011;12: 745–55. doi:10.1038/nrg3031

9. Curriculum vitae

Persönliche Daten:

Name:	Thomas Schmutzer
Geburtsdatum:	24.06.1982
Geburtsort:	Leipzig (Sachsen)
Email:	schmutzr@ipk-gatersleben.de

Wissenschaftlicher Werdegang:

03/2009 - heute	**Wiss. Mitarbeiter** am Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) in Gatersleben
06/2012 - 11/2012	**Wiss. Mitarbeiter** (Forschungsaufenthalt) am INRA in Clermont-Ferrand Vergleichende genomische Analyse des Weizenchromosoms 3B und des Gerstechromosoms 3H
04/2008 - 12/2008	**Diplomand** am IPK in Gatersleben Thema der Diplomarbeit: „Evaluierung von Assemblierungswerkzeugen für Hochdurchsatz-Sequenzierungs-Techniken“
10/2003-02/2009	**Studium** der Bioinformatik an der Martin-Luther-Universität Halle-Wittenberg

---

Thomas Schmutzer

Jena, 25.05.2016

10. Further publications

The following list of publications is given in alphabetical order of first author. The four publications embedded in this cumulative thesis are not listed.

Ariyadasa, R., Mascher, M., Nussbaumer, T., Schulte, D., Frenkel, Z., **Schmutzer, T.**, … Stein, N. (2014). A sequence-ready physical map of barley anchored genetically by two million single-nucleotide polymorphisms. *Plant Physiology*, *164*(1), 412–23. doi:10.1104/pp.113.228213

Cao HX, **Schmutzer T**, Scholz U, Pecinka A, Schubert I, Thi G, et al. Metatranscriptome analysis reveals host-microbiome interactions in traps of carnivorous Genlisea species. *Frontiers in Microbiology* 2015; doi:10.3389/fmicb.2015.00526

Colmsee C, Beier S, Himmelbach A, **Schmutzer T**, Stein N, Scholz U, et al. BARLEX – the Barley Draft Genome Explorer. *Mol Plant.* 2015; 1–3. doi:10.1016/j.molp.2015.03.009

Kohl, S., Hollmann, J., Blattner, F. R., Radchuk, V., Andersch, F., **Schmutzer, T.**, … Weschke, W. (2012). A putative role for amino acid permeases in sink-source communication of barley tissues uncovered by RNA-seq. *BMC Plant Biology*, *12*(1), 154. doi:10.1186/1471-2229-12-154

Martis, M. M., Klemme, S., Banaei-Moghaddam, A. M., Blattner, F. R., Macas, J., **Schmutzer, T**., … Houben, A. (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *PNAS*, *109*(33), 13343–6. doi:10.1073/pnas.1204237109

Martis, M. M., Zhou, R., Haseneyer, G., **Schmutzer, T**., Vrána, J., Kubaláková, M., … Stein, N. (2013). Reticulate evolution of the rye genome. *The Plant Cell*, *25*(10), 3685–98. doi:10.1105/tpc.113.114553

Pellino, M., Hojsgaard, D., **Schmutzer, T**., Scholz, U., Hörandl, E., Vogel, H., & Sharbel, T. F. (2013). Asexual genome evolution in the apomictic Ranunculus auricomus complex: examining the effects of hybridization and mutation accumulation. *Molecular Ecology*. doi:10.1111/mec.12533

Poursarebani, N., Ma, L., **Schmutzer, T**., Houben, A., & Stein, N. (2014). FISH Mapping for Physical Map Improvement in the Large Genome of Barley : A Case Study on Chromosome 2H. *Cytogenetic and Genome Research*. doi:10.1159/000366028

Snowdon RJ, Abbadi A, Kox T, **Schmutzer T**, Leckband G. Heterotic Haplotype Capture: precision breeding for hybrid performance. *Trends Plant Science* 2015;20: 410–413. doi:http://dx.doi.org/10.1016/j.tplants.2015.04.013

Taudien, S., Steuernagel, B., Ariyadasa, R., Schulte, D., **Schmutzer, T**., Groth, M., … Platzer, M. (2011). Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Research Notes*, *4*(1), 411. doi:10.1186/1756-0500-4-411

11. Acknowledgments

Bei Matthias Platzer möchte ich mich für die Ermöglichung der Anfertigung meiner Dissertation unter seiner Betreuung bedanken. Das entgegengebrachte Interesse an meinen Arbeiten und die zielführenden Diskussionen waren von großer Bedeutung für die erfolgreiche Anfertigung der Dissertation. Ein besonderer Dank gilt weiterhin meinem Arbeitsgruppenleiter Uwe Scholz, der mit Vertrauen, Förderung und Forderung von Selbständigkeit meine beständige Weiterentwicklung prägte. Ebenso gilt Nils Stein mein Dank, der stets eine hervorragende wissenschaftliche Atmosphäre geschaffen und mich insbesondere gelehrt hat, selbstkritisch jedes Ergebnis mehrfach zu hinterfragen. Des Weiteren ist es mir wichtig, allen Koautoren und Begleitern zu den wissenschaftliche Publikationen zu danken, ohne die diese Dissertation nicht angefertigt hätte werden können. Besonders hervorzuheben ist hierbei Eva Bauer als Projektkoordinatorin des RYE-Express und des RYE-Select Projekts. Mit ihrer langjährigen Förderung der Roggenforschung hat sie auch für meine wissenschaftliche Entwicklung eine wichtige Rolle eingenommen.

Der gesamten Arbeitsgruppe BIT und damit allen jetzigen und ehemaligen Arbeitskollegen danke ich für ein so sympathisches Umfeld. Wissenschaft macht Spaß und das auf vielen Ebenen. Burkhard, als ehemaliger Kollege, gilt mein Dank für das gemeinsames Vordringen in das Gebiet der Genomanalysen und die damit verbundene Bewältigung zahlreicher Herausforderungen, vor allem zu Beginn meiner Zeit am IPK. Zu erwähnen ist auch, dass eine wichtige Basis meiner tagtäglichen Arbeit durch die hervorragende IT-Unterstützung gegeben war. Dafür gilt Steffen, Thomas M., Jens und Heiko ein großes Dankeschön. Des Weiteren möchte ich Fabian und Chris erwähnen, die als fleißige Studenten eine große Unterstützung waren.

Für das Korrekturlesen meiner Dissertation danke ich Doreen, Natasha, Matthew, Uwe, Martin, Sebastian, Christian, Bernd, Zhou und Simone. Euer Feedback und die intensive Auseinandersetzung waren eine sehr wichtige Unterstützung.

Zum Abschuss möchte ich meiner Familie und all den Menschen danken, die mich ein Leben lang gefördert und auch während der Erstellung der Dissertation beständig unterstützt haben. Mein herzlichster Dank gilt hierbei Simone. Deine große Geduld, Rücksicht und Motivation waren enorm wichtig. Deine Unterstützung hat mir die Kraft gegeben auch diese Herausforderung zu bestehen.

Eigenständigkeitserklärung

Entsprechend der geltenden, mir bekannten Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Friedrich-Schiller-Universität Jena erkläre ich, dass ich die vorliegende Dissertation mit dem Titel „Strategies to Detect Genetic Diversity in Plants“ eigenständig angefertigt und alle von mir benutzten Hilfsmittel und Quellen angegeben habe. Es wurde weder die Hilfe eines Promotionsberaters in Anspruch genommen, noch haben Dritte für Arbeiten, welche im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen, unmittelbare oder mittelbare geldwerte Leistungen erhalten.

Die vorgelegte Dissertation wurde noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Weiterhin habe ich mich mit der vorliegenden Arbeit an keiner anderen Hochschule um den akademischen Grad doctor rerum naturalium (Dr. rer. nat.) beworben und weder früher noch gegenwärtig die Eröffnung eines Verfahrens zum Erwerb des o.g. akademischen Grades an einer anderen Hochschule beantragt.

Thomas Schmutzer

Jena, 25.05.2016